RATTLESNAKE GENOMICS ILLUSTRATE PATTERNS OF SPECIATION, ADAPTATION, AND LINKS BETWEEN GENOME STRUCTURE AND FUNCTION

by

DREW ROBERTS SCHIELD

Presented to the Faculty of the Graduate School of

The University of Texas at Arlington in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON

MAY 2018

Acknowledgements

I am grateful to a number of people who have made life and work more enjoyable and productive during my dissertation. First, thank you to my friends and lab mates in the Castoe Lab – Rich Adams, Blair Perry, Giulia Pasquesi, Nicky Hales, Andrew Corbin, Audra Andrew, Jacobo Reyes-Velasco, and especially Daren Card, who started in the lab at the same time as me, and with whom I've been a package deal ever since (still holding out for a 'spousal' hire down the road). You guys have been a joy to work with, and represent an enormous source of support and commaraderie. I am grateful to also work with wonderful collaborators outside of the lab; thank you to Drs. Steve Mackessy, Jesse Meik, Tereza Jezkova, Heath Blackmon and your labs. Thank you to all of my friends and colleagues in the UTA biology department, especially to Shannon Beston, James Titus MacQuillan, and Kyle Shaney. Thank you to the Smith, Fujita, Demuth, Betran, Chippendale, and Walsh labs for sharing ideas and providing unique perspectives. Thank you also to my committee members, Drs. Esther Betran, Jeff Demuth, Matt Fujita, and Matt Walsh, for your support, knowledge, and ideas for improving my research. A huge thank you to the biology department staff – to Rachel Wostl, Linda Taylor, Gloria Burlingham, and Ashley Priest, thank you for putting up with me! I am especially grateful for the tireless support of my family (Mom, Dad, Aubrey), Darren Baun, Caleb Rasmussen, Sally Rabun and the Rabun Family, The Deluxe (Justin Stettner, Uriel Avila, and Jesse Meik; he's on here twice!), and Jill Castoe. Lastly, thank you to my advisor, Todd Castoe, for being an outstanding mentor, orator of famous 'Rocky' pep talks, and friend for the last five and half years – I'm eager to see what we'll cook up next; "All of your snakes are belong to us."

February 16th, 2018

Dedication

This dissertation is dedicated to my parents, Donavon and Karen Schield. To my mother, for teaching me to be curious about nature and encouraging my interests in science and evolution, especially the biology of snakes and all manner of crawling creatures. To my father, for teaching me the importance of focus, being industrious, and developing a strong work ethic, so that I might meaningfully pursue them.

Abstract

RATTLESNAKE GENOMICS ILLUSTRATE PATTERNS OF SPECIATION, ADAPTATION, AND LINKS BETWEEN GENOME STRUCTURE AND FUNCTION

Drew R. Schield, PhD

The University of Texas at Arlington, 2018

Supervising Professor: Todd A. Castoe, PhD

Understanding the origins of species and biological novelties that allow them to thrive in diverse environments is a key goal in evolutionary biology, and new genomic methods are constantly enabling research using non-model species to address important questions related to speciation and adaptation. Using phylogeographic, population genetic, and comparative genomic methods, I demonstrate that North American rattlesnakes are a uniquely enriched system for investigating patterns and processes at the intersection of adaptation and speciation. Specifically, this dissertation explores the evolution of biological novelty at multiple scales, including the origins of reproductive incompatibilities during the process of gene flow in secondary contact, evidence for links between genomic patterns of selection and locally adapted traits (e.g., venom and reproductive phenotypes), and cryptic genetic diversity in widely-distributed rattlesnake lineages. Detailed investigation of the high-quality prairie rattlesnake genome provides new perspectives into the evolution of genome structure in vertebrates, sex chromosome differentiation, the unique biology and significance of microchromosomes, and the origins of venom, one of the most distinctive features of rattlesnake biology. Collectively, this work serves as an example of the tremendous value that rattlesnakes hold for addressing important evolutionary questions in the age of genomics.

Table of Contents

Chapter 1

Introduction

Widely-distributed species complexes can span a broad array of ecologies, habitats, and ecological niches, and thus are valuable study systems capable of yielding new information regarding diversification patterns and speciation because they provide unique opportunities to observe the recent and ongoing influence of selection and gene flow on these processes. Closely related lineages within these complexes that occupy diverse habitats and show contrasting phenotypes can inform us about the evolution of ecologically relevant genotypic and phenotypic novelty that stems from genomic variation in otherwise highly similar genetic backgrounds. In the age of genomics, we have been increasingly able to understand the links between genotypic and phenotypic variation in 'non-model' systems (i.e., organisms other than *Drosophila* and mice, for example), which only serves to provide more numerous and exciting examples of the genomic underpinnings of biological innovations in nature.

My dissertation has focused on one such system, a group of widely-distributed North American rattlesnakes native to the western United States and Mexico. The the focal species of this work (the western diamondback rattlesnake, *Crotalus atrox*; the Mojave rattlesnake, *Crotalus scutulatus*; and the western rattlesnake species complex, *Crotalus viridis*, *oreganus*, and other subspecies) diverged from a common ancestor roughly 6-8 million years ago, and have since grown to occupy a rich diversity of ecological regions including deserts and grasslands, and occur at a range of elevations and climatic conditions. Their evolutionary history has also been shaped by climatic flucations during the Pliocene and Pleistocene and shifts in the geography of suitable habitat, which has driven prolonged periods of divergence in isolation followed by introgression in secondary contact. They also exhibit remarkable variation in size, coloration,

and most notably, their venom composition. The diversity of their venoms alone, and the hypothesis that geographic venom variation is due to local adaptation to available prey, has driven analyses in several chapters of this dissertation, which aim to better understand the nature of venom variation in hybridization of divergent lineages (Chapter 3) and the genomic structure and regulation of the venom phenotype (Chapter 5). Other chapters of this dissertation have sought to develop a ground work understanding of the population structure and gene flow dynamics among populations across the entire range of species (Chapters 2 and 4) for further detailed studies of speciation and adaptation, and have identified evidence of cryptic diversity punctuated by geographic and ecological barriers to gene flow (Chapter 4).

The techniques and foci of the following four chapters are wide in scale and scope, ranging from detailed examinations of the roles of selection in divergence and introgression (Chapter 3), to broader scale analyses of phylogeny, phylogeography, population genetic structure, and gene flow (Chapters 2 and 4), to comparative genomics among reptiles and amniotes (Chapter 5); these studies have yielded exciting findings with implications for our understandings of speciation, adaptation, and the evolution of genomic structure and function.

Ultimately, this dissertation will hopefully serve as an example of the tremendous potential that rattlesnakes hold has a system for understanding a diverse array of evolutionary patterns and processes, and also one that raises as many questions as it answers, which will lead to timely and impactful extensions of this work. There is still a great deal to learn about the intriguing genomic biology of these dynamic and charasmatic rattlesnakes, and the approaches and resources established by this work should serve as a foundation for further research to leverage this system, which is truly enriched for studying adaptation and speciation in the age of genomics.

Chapter 2

Incipient speciation with biased gene flow between two lineages of the Western Diamondback Rattlesnake (*Crotalus atrox*)

Drew R. Schield[a], Daren C. Card[a], Richard H. Adams[a], Tereza Jezkova[b], Jacobo Reyes-Velasco[a], F. Nicole Proctor[a], Carol L. Spencer[c], Hans-Werner Herrmann[d], Stephen P. Mackessy[e], and Todd A. Castoe[a,*]

[a]Department of Biology & Amphibian and Reptile Diversity Research Center, 501 S. Nedderman Drive, University of Texas at Arlington, Arlington, TX 76019 USA

[b]School of Life Sciences, University of Nevada, Las Vegas, 4505 Maryland Parkway, Las Vegas, Nevada, 89154 USA

[c]Museum of Vertebrate Zoology, 3101 Valley Life Sciences Building, University of California, Berkeley, CA 94720 USA

[d]School of Natural Resources and the Environment, 1041 E Lowell Street, University of Arizona, Tuscon, AZ 85721 USA

[e]School of Biological Sciences, 501 20th Street, University of Northern Colorado, Greeley, CO 80639 USA

Abstract

We used mitochondrial DNA sequence data from 151 individuals to estimate population genetic structure across the range of the Western Diamondback Rattlesnake (*Crotalus atrox*), a widely distributed North American pitviper. We also tested hypotheses of population structure using double-digest restriction site associated DNA (ddRADseq) data, incorporating thousands of nuclear genome-wide SNPs from 42 individuals. We found strong mitochondrial support for a deep divergence between eastern and western *C. atrox* populations, and subsequent intermixing of these populations in the Inter-Pecos region of the United States and Mexico. Our nuclear RADseq data also identify these two distinct lineages of *C. atrox*, and provide evidence for nuclear admixture of eastern and western alleles across a broad geographic region. We identified contrasting patterns of mitochondrial and nuclear genetic variation across this genetic fusion zone that indicate partially restricted patterns of gene flow, which may be due to either pre- or post-zygotic isolating mechanisms. The failure of these two lineages to maintain complete genetic isolation, and evidence for partially-restricted gene flow, imply that these lineages were in the early stages of speciation prior to secondary contact.

Introduction

Speciation proceeds with the origin of barriers to gene flow, which permits the maintenance of genetic and phenotypic divergence (Coyne and Orr 2004; Nosil and Feder 2012). Isolation mechanisms may vary in strength, leading to a continuum of speciation based on the degree to which such mechanisms promote reproductive isolation between lineages (Coyne and Orr 2004; Nosil and Feder 2012). Lineages in the early stages of this speciation continuum are particularly valuable as model systems for understanding the mechanisms that drive speciation. They are valuable for understanding primary mechanisms of isolation because they lack confounding secondary isolation mechanisms that evolve later (Orr 1995; Good et al. 2008). Accordingly, studying these model systems may also provide novel insight into why speciation might not occur, leading to the secondary fusion of divergent lineages (Taylor et al. 2006; Wiens et al. 2006; Webb et al. 2011). Comparing patterns of mitochondrial and nuclear genetic variation can be useful for identifying such patterns of partial genetic isolation, due to their different modes of inheritance. Here we conduct a detailed analysis of nuclear and mitochondrial gene flow between two divergent lineages of the Western Diamondback Rattlesnake (*Crotalus atrox*), and assess evidence for where these divergent lineages may exist on the speciation continuum.

The Western Diamondback Rattlesnake is a large venomous rattlesnake native to the United States and Mexico. It inhabits among the broadest distributions of all rattlesnake species, ranging across much of the southwestern United States and northern Mexico (Campbell and Lamar 2004). This species is viewed as a habitat and diet generalist, living in a diversity of lowland habitats and feeding on a large variety of prey. There are no described subspecies within *C. atrox*, although there is evidence of morphological variation across their range (Klauber 1956; Spencer 2008). Given its large range, local abundance, large size, and toxic venom, *C. atrox* is

among the most medically relevant species in North America, with the highest number of human fatalities due to envenomation of any snake (Campbell and Lamar 2004).

In a previous study, Castoe et al. (2007) found evidence for two distinct *C. atrox* mitochondrial lineages that diverged approximately 1.36 MYA, corresponding roughly to populations east and west of the Continental Divide of the United States and Mexico. Although sampling was limited, the authors found some evidence that these two mitochondrial lineages overlap in distribution near the Continental Divide, suggesting possible gene flow between these lineages. Although mitochondrially-encoded loci have been used extensively to study snake phylogeography and population genetics (e.g., (Ashton and de Queiroz 2001; Douglas et al. 2002; Castoe et al. 2007; Burbrink and Castoe 2009; Meik et al. 2012), they may obscure some patterns of introgression due to their matrilineal inheritance, lack of recombination, and rapid coalescence (Avise and Vrijenhoek 1987; Palumbi and Baker 1994; Avise 2000). The combination of mitochondrial data with data from nuclear single nucleotide polymorphisms (SNPs), however, provides a more powerful means of testing population genetic hypotheses and of comparing matrilineal versus whole genome patterns of genetic variation.

Here we combined extensive sampling of individuals for a mitochondrial gene with sampling of thousands of nuclear SNPs from restriction site associated DNA sequencing (RADseq) to investigate patterns of historical and current divergence, and gene flow across the range of *C. atrox*. Our aim in this study was to use both mitochondrial and nuclear data to characterize patterns of divergence and secondary contact in *C. atrox*, and to detect signatures of progression towards speciation of lineages within *C. atrox* prior to recent introgression. To address this aim, we tested the following hypotheses: 1) Mitochondrial and nuclear data provide evidence of two divergent *C. atrox* lineages, 2) Mitochondrial and nuclear data show evidence of recent gene

flow between these lineages, 3) Both eastern and western populations show similar patterns of genetic diversity and historical demography based on mitochondrial and nuclear data, and 4) Gene flow is unrestricted between eastern and western lineages, and there is no evidence of reproductive isolation.

Materials and Methods

Taxon sampling and DNA extraction

We obtained tissues from a total of 151 *Crotalus atrox* from throughout their range (Fig. 1B; Table S1), including the 48 samples in Castoe et al. (2007). Tissues included samples of blood, liver, and skin preserved by snap-freezing or via lysis buffer, RNALater, shed skins or rattles. Genomic DNA was isolated in one of four ways: using a Qiagen DNeasy extraction kit (samples from Castoe et al. 2007; Qiagen, Inc., Valencia, CA, USA), Zymo Research Genomic DNA Tissue MiniPrep kit (most solid tissues; Zymo Research Corporation, Irvine, CA, USA), Thermo Scientific GeneJet whole blood DNA extraction kit (blood; Thermo Fisher Scientific, Inc., Waltham, MA, USA), or phenol-choroform-isoamyl alcohol (some shed skins).

Mitochondrial locus amplification and sequencing

We used PCR to amplify a fragment of the mitochondrially-encoded NADH dehydrogenase subunit 4 (ND4), plus the downstream Serine, Histidine, and Leucine tRNAs, using the primers ND4 and Leu (Arevalo et al. 1994). PCR products were purified using AgenCourt AMPure XP beads (Beckman Coulter, Inc., Irving, TX, USA). Purified PCR products were quantified and sequenced in both directions with the use of amplification primers, using BigDye on an ABI 3730 capillary sequencer (Life Technologies, Grand Island, NY, USA) at the UTA Genomics Core Facility.

ddRADseq library generation and sequencing

A subset of the DNA samples used for ND4 PCR were also used to generate ddRAD libraries (n = 42; Fig. 3B, Supplementary Online Table 1), largely following the protocol of Peterson et al. (2012) except for notable exceptions outlined below. Samples were chosen based on locality in order to include representative specimens of the putative eastern, western, and hybrid zone populations. Genomic DNA was digested using a combination of rare (*SbfI*; 8bp recognition site) and common (*Sau3AI*; 4bp recognition site) cutting restriction enzymes. Double-stranded indexed DNA adapters were ligated to the ends of digested fragments that also contained unique molecular identifiers (UMIs; eight consecutive N's prior to the ligation site). Following adapter ligation, samples were pooled in sets of eight, size selected for a range of 590-640 bp using the Blue Pippin Prep (Sage Science, Beverly, MA, USA), and PCR amplified using primers to complete attachment of flow-cell binding sequences and addition of a second index specific to each sub-pool. Sub-pools were pooled again based on molarity calculations from analysis on a Bioanalyzer (Agilent, Santa Clara, CA, USA) using a DNA 7500 chip, and sequenced using 100 bp paired-end reads on an Illumina HiSeq 2500.

mtDNA sequence analysis

Raw mitochondrial gene sequence chromatograms were edited using Geneious v6.1.6 (Biomatters Ltd., Auckland, NZ), and we aligned edited sequences using MUSCLE (Edgar 2004) with minimal manual adjustments to improve alignment of tRNA gene regions and to trim the 5' and 3' ends of all sequences to reduce columns with high levels of missing data. The final alignment contained 813 aligned bases and included no indels.

We estimated phylogenetic relationships among unique *C. atrox* haplotypes and outgroup species using Bayesian phylogenetic inference in MrBayes v3.2.1. (Huelsenbeck and Ronquist 2001).

For outgroups, we used single representatives *C. molossus, C. tigris, and C. ruber*, which were chosen based on previous estimates of relationships among rattlesnakes (Reyes-Velasco et al. 2013) We used the Bayesian Information Criterion (BIC) implemented in PartitionFinder v1.1.1 (Lanfear et al. 2012) to select best-fit models, which were HKY for 1st and 2nd codon positions, as well as for tRNAs, and TN + Γ + invariant sites for 3rd codon positions. This partitioned model was used for analyses in MrBayes, which consisted of four runs, each run for 10^7 generations with four chains (one cold and three heated), sampled every 500 generations. We accounted for among partition rate variation using the "prset ratepr = variable" command. Potential scale reduction factor value estimates (PSRF) indicated that individual runs had converged by 10^5 generations, and thus we discarded the first 10^5 samples as burn-in. We also confirmed that independent runs had converged based on overlap in likelihood and parameter estimates among runs, as well as effective sample size (ESS) and PSRF values, which we evaluated in Tracer v1.5 (Drummond and Rambaut 2007). We generated a 50% majority rule consensus phylogram using combined estimates from post burn-in samples from independent runs. We also constructed a median-joining haplotype network to visualize relationships among unique haplotypes using Network v4.5.1.6 (Bandelt et al. 1999), weighting transitions 2:1 over transversions (recommended in the Network manual to distinguish between haplotype connections that would be equally parsimonious with a 1:1 ratio) and using the maximum parsimony option to reduce excess links among haplotypes from the resulting network. We estimated haplotype diversity in each population using the Nei and Tajima equation (Nei and Tajima 1981) implemented in custom Python scripts.

We analyzed changes in effective population size through time in the two primary mitochondrial lineages of *C. atrox* using the Bayesian Skyline Plot (BSP) coalescent model (Drummond et al.

2005) implemented in BEAST v1.7.5 (Drummond and Rambaut 2007). We partitioned the dataset by gene (ND4 and tRNA) and ND4 was further partitioned by codon position (1st and 2nd position combined, 3rd position) and used the substitution model HKY, along with a strict molecular clock (recommended for intraspecific inferences in the BEAST manual) and a coalescent Bayesian skyline tree prior. We applied a 0.7% per lineage per million years mutation rate, which has been used in other studies using the ND4 gene in snakes, including *C. atrox* (Wuster et al. 2002; Castoe et al. 2007; Lane and Shine 2011b). We conducted two independent runs of 4 x 10^7 generations with a burn-in of 25%. Additionally, we used IMa2 (Hey and Nielsen 2007) to obtain estimates of the marginal posterior probability density of the parameters of the isolation-migration model (Hey and Nielsen 2004), including t_0 (time since population splitting), q_0 (effective western population size), q_1 (effective eastern population size), q_A (effective ancestral population size), $m_1 > m_0$ (migration rate from west to east), and $m_1 < m_0$ (migration rate from east to west). We ran IMa2 using our total mtDNA dataset, with population sample sizes assigned based on the locality of each individual relative to the Continental Divide. Each of four independent runs had a 2 x 10^6 generation burn-in period (length was determined in initial trial runs) followed by a 10^7 generation post-burn-in sampling period, which we determined based on chain mixing and ESS values exceeding 1000 for parameters in all runs, and convergence was assessed by congruent results from independent runs. We rescaled parameter estimates using generation time and mutation rate estimates for *C. atrox* from Castoe et al. (2007).

We assessed landscape-level patterns of genetic differentiation for the eastern and western mitochondrial clades separately by interpolating pairwise mitochondrial genetic distances (mismatch distances) among sampling localities. Mismatch distances were calculated in Alleles in Space v1.0 (Miller 2005) and assigned to geographic coordinates representing midpoints

between sampling localities using the Delaunay triangulation-based connectivity network (Miller et al. 2006). To account for the correlation between genetic and geographical distances, we used residual values derived from the linear regression of mismatch distances derived from the linear regression of mismatch versus geographical distance (Manni et al. 2004). Any data points with residuals outside of the 95% confidence interval of the linear regression were removed from analysis. We imported the mismatch distances and associated coordinates into ArcGIS v9.2 (ESRI, Redlands, CA, USA) and interpolated them across 2.5-minute grids using the inverse distance weighted procedure (Watson and Philip 1985) in the ArcGIS Spatial Analyst extension. In order to restrict the interpolation analysis within an area that is actually occupied by the species, we approximated the range of *C. atrox* via ecological niche modeling implemented in MAXENT v3.2.1 (Phillips et al. 2006) using default parameters, with 50 replicates per population, and average model probabilities converted to presence-absence maps in ArcGIS. This methodology extracts environmental data (obtained from the Worldclim dataset; (Hijmans et al. 2005) corresponding to occurrence records which were represented by geographic coordinates obtained from HerpNet (www.herpnet.org) and evaluates habitat suitability across the landscape using program-specific algorithms (Elith et al. 2006). The resulting maps were used as an approximation of the range of *C. atrox* for geographically masking the surface derived from the interpolation analyses.

To estimate how cooler climatic cycles during the Pleistocene might have restricted the range of *C. atrox*, we estimated the geographic distribution of the eastern and the western population during the last glacial maximum (LGM). We projected the present-day models for eastern and western populations of *C. atrox* (discussed above) onto climatic reconstructions of the LGM under the assumption that the climatic niche of each population remained conserved between the

LGM and present time (Elith et al. 2010; Jezkova et al. 2011). For environmental layers representing the climatic conditions of the LGM, we used ocean-atmosphere simulations available through the Paleoclimatic Modelling Intercomparison Project (Braconnot et al. 2007). We used Community Climate System Model v. 3 (CCSM) that has been previously downscaled to the spatial resolution of 2.5 minutes and converted to bioclimatic variables (Waltari et al. 2007). We constructed LGM models in MAXENT using default parameters and ran models for 20 replicates per model, and obtained an average model using logistic probability classes of climatic niche suitability. Maps representing suitable climatic niche for eastern and western populations were determined using a logistic probability threshold that balances omission, predicted area, and a threshold value (Liu et al. 2005).

Analysis of RADseq data

Raw ddRADseq Illumina sequencing reads were processed using the Stacks pipeline (Catchen et al. 2013). PCR clones were removed using the Stacks clone filter program using the in-line UMI regions of our adapter design, which were subsequently trimmed away using the FASTX-Toolkit trimmer (Hannon 2014). Trimmed reads were processed using the process radtags function in Stacks, which parses reads by barcodes, confirms the presence of restriction digest cut sites, and discards reads that lacked these features or those with poor quality scores. We used the *de novo* main Stacks pipeline, including Ustacks, Cstacks, and Sstacks, to summarize SNP information for downstream analyses.

We used the populations program in Stacks to obtain various population genetic parameters estimates. We used two alternative population assignments for analyses: 1) all individuals as a single population (these outputs were used for downstream Structure analyses); and 2) a two-population model, with individuals partitioned into western and eastern populations. For the two-

population model, we only used individuals that fit the following two criteria: 1) nuclear allele assignments of >90% to either the western or eastern population cluster in Structure analyses, and 2) mitochondrial haplotypes that matched with the majority of nuclear genome assignments (e.g., >90% western nuclear alleles and containing a western mitochondrial haplotype); this population assignment model was used for genetic diversity comparisons between western and eastern populations. We used thresholds for missing data (50%) as well as minimum read depth per stack (5) in populations for all analyses. We used these thresholds to maximize the number of loci available for analyses after determining that other threshold settings (e.g., 75% missing data, 10X stack depth) did not qualitatively alter our estimates (Supplementary Online Table 2).

We inferred population structure and admixture using Structure (Pritchard et al. 2000), based on 4,494 SNPs recovered by our Stacks analyses. We first estimated the allele frequency distribution parameter (λ), using a trial run with *K* set to 1. We used this estimation (λ = 3.13) in short clustering runs (10^4 burn-in, 10^4 data collection) under a mixed ancestry model (Hubisz et al. 2009) for *K* = 2-9 with no putative population origins specified for individuals (single population model). From this, we determined that a *K* range of 2-5 population clusters was most likely based on model likelihood estimates. We then performed longer runs (10^7 burn-in, 10^8 data collection) for *K* = 2-5 (each with 3 iterations). For each of these runs, the optimal *K* clusters was determined using the Δ*K* method described by Evanno et al. (2005) implemented in StructureHarvester (Earl and Vonholdt 2012). The results of Structure runs were visualized using Distruct (Rosenberg 2004).

To quantify genomic introgression in *C. atrox*, we conducted Bayesian estimation of genomic clines using the program bgc (Gompert and Buerkle 2011). This program estimates the

probability that an individual with a hybrid index *H* has inherited a gene copy at a given locus from one of two parental populations using two locus-specific cline parameters (α, the genomic cline center parameter, and β, the genomic cline rate parameter; estimated using MCMC). Under this model, if α and β both equal zero, an individual's hybrid index will be equal to the probability of ancestry (ϕ) from a given parental population (Gompert and Buerkle 2011; Gompert et al. 2012b). We were specifically interested in genome wide estimates of both cline parameters to understand the degree of variation in locus specific introgression across the genome of our admixed population. We used data from all individuals included in our RADseq dataset, and partitioned samples into western parental, eastern parental, and admixed populations. We used allele output files for each individual from Stacks and custom Python scripts to generate biallelic input files for each population, which included information for a total of 19,123 loci (though some loci suffered from missing data for particular individuals). We ran bgc using the genotype uncertainty model (recommended for next-generation sequencing data; see bgc manual) on our parental and admixed population dataset using four chains for 50,000 generations each, discarding the first 5,000 generations as burnin, and using default settings assuming free recombination between all loci (maximum distance between loci set to 0.5). We then combined the output from the four chains after inspecting the MCMC output for convergence onto a stationary distribution.

Results

Mitochondrial gene variation

Within *C. atrox*, we identified 52 unique ND4 haplotypes in our sampling. The four independent runs of Bayesian phylogenetic analyses using unique haplotypes converged on nearly identical estimates of the likelihood score, had PSRF values very near 1.0 throughout non burn-in

generations, and had ESS values above 400 for all parameters. The 50% majority rule consensus phylogram from these runs is presented in Fig. 1A. Within *C. atrox,* we found strong support (>0.95 posterior probability) for a deep split between two haplotype lineages that correspond approximately to populations east and west of the Continental Divide of the US and Mexico (Fig. 1B). Both of these major clades, however, contain subclades that include haplotypes from individuals on either side of the Continental Divide. The two major mitochondrial clades are also apparent in the median-joining haplotype network (Fig. 1C). The eastern clade contains a greater number of haplotypes (31 of 52 haplotypes) compared to the western clade. Despite this, the western population has a higher degree of haplotype diversity (0.89) than the east (0.75), which appears to be due to many eastern samples possessing one of a small number of high-frequency haplotypes.

Our map of eastern and western haplotypes highlights a putative zone of population fusion, which we term the Inter-Pecos region, implicating the Continental Divide as a barrier to gene flow for eastern haplotypes, as very few individuals belonging to this clade were observed west of the divide (Fig. 1B). Likewise, we do not observe any western haplotypes east of the Pecos River in Texas, suggesting that this region in general represents a barrier to gene flow for the western clade. The results of our BSP analysis of historical demography in eastern and western clades indicate that both lineages have experienced recent population expansion (i.e., within the last 100 KYA). These results also suggest that the eastern population has undergone a greater increase in effective population size relative to the west (approximately 1.7 million and 3 million in the west and east, respectively; Supplementary Online Figure 1).

Our independent IMa2 runs resulted in nearly identical estimates of marginal posterior probability densities for each parameter, and ESS values were greater than 2,000 for all

parameters; we present the results from the independent run with the highest ESS values in Supplementary Online Figure 2. Based on our IMa2 analyses, we found evidence of substantial population expansion in both the eastern and western population (Supplementary Online Figure 2A), with a greater degree of population expansion in the eastern population; these findings are all consistent with our inference from BSP analysis (Supplementary Online Figure 1). We were unable to obtain a robust estimate of the ancestral effective population size from IMa2 analyses, however, as parameter estimation appears to have suffered from insufficient data resulting an essentially flat posterior distribution for this parameter. The posterior estimates for migration rate parameters support very low migration rates (~7 x 10^{-8} migrants per year) for both populations (Supplementary Online Figure 2B), though estimates for these parameters may have also suffered from insufficient data for accurate estimation of migration rates as indicated by posterior densities that do not reach low levels near the upper or lower limits of the prior (even after multiple runs with increasingly broad priors). We found support for a TMRCA for the eastern and western populations of approximately 675,000 years ago (95% posterior interval values ranging from 412,933 - 1,027,939; Supplementary Online Figure 2C), which falls well within the time span of Pleistocene glaciation cycles.

When residual pairwise genetic distances are interpolated on geographic distance and are visualized across the range of both clades, we find evidence for highly contrasting levels of genetic diversity. Although there are varying degrees of diversity in each lineage, there is an overall greater degree of genetic diversity (inferred via mismatch distances) in the east (range = -0.012 – 0.029; Fig. 2B) relative to the west (range = -0.012 – 0.012; Fig. 2A), particularly in the central and western regions of Texas. Both lineages harbor relatively low diversity within areas of the Inter-Pecos fusion zone, consistent with range expansion in each lineage towards the

Continental Divide. Furthermore, we find that a large portion of the high-diversity core observed in the eastern lineage falls within the eastern portion of the Inter-Pecos region, while the western lineage shows no evidence of high diversity anywhere within the Inter-Pecos.

The results of our LGM climatic niche modeling highlight major differences in estimated Pleistocene refugia for the eastern and western populations, and provide context for our observed patterns of contrasting diversity in western versus eastern populations (Fig. 2C). At all thresholds, we find that the eastern population inhabited a large region of the Chihuahuan Desert and adjacent Gulf Coastal Plains. This region is comprised of multiple large segments that are largely in contact; thus, the eastern refugium exhibits little fragmentation, whereas the western refugium occupies a much smaller and more linear range. Our models also demonstrate that the eastern population occupied part of the Inter-Pecos region during the LGM, while the western population was completely absent. Additionally, these models suggest that the eastern and western populations of *C. atrox* were most likely not in contact during Pleistocene glacial cycles, and were instead isolated by an extensive region of high-elevation.

Nuclear SNP variation

RADseq filtering thresholds of 50% missing data per locus and 5X read depth per locus provided the most numerous RAD loci (4,519) for analyses among threshold combinations used under the two-population model explained above, and genetic diversity estimates did not vary qualitatively with this filtering scheme relative to others (for results from Stacks populations analyses under different filtering thresholds, see Supplementary Online Table 2). We find substantial population differentiation among western and eastern populations based on these data (F_{ST} = 0.15). This estimate is consistent with moderate-to-high levels of differentiation (Lewontin 1972; Lewontin

and Krakauer 1973), and emphasizes that, prior to secondary contact, eastern and western lineages were well-differentiated incipient species.

Our nuclear SNP dataset provided consistent evidence of higher genetic diversity in the eastern population relative to the west. We find a much higher number of private alleles in the eastern population than in the western population (453 and 253, respectively; Fig. 3A). Observed heterozygosity is significantly higher in the eastern population than in the western population ($p < 0.0001$; Fig. 3B). We find the same pattern in the estimates of nucleotide diversity (π), with significantly higher diversity in the east ($p < 0.0001$; Fig. 3C). Altogether, the pattern of higher nuclear genetic diversity in the eastern population is consistent with similar patterns observed in our mitochondrial dataset.

In our final Structure analyses, we estimated $K = 4$ as the optimal model of population clustering, and populations assignments are shown in Figure 3D. Additionally, we visualized the results of the two-cluster ($K = 2$) model for comparison, because a two-population model was most consistent with the general findings from our mitochondrial data. We find that both two and four population models recover similar population structure, such that there are identifiable western and eastern clusters. Both models also indicate a similar fusion zone between lineages, with the majority of individuals within the zone showing evidence of substantial admixture. These samples predominately fall within the Inter-Pecos region as defined by our mitochondrial dataset, although nuclear SNPs highlight the Continental Divide as the location of a very steep cline of genotype intergradation between eastern and western population clusters (also see section 3.3. below).

When we ordered samples by longitude (as shown in Fig. 3D), we found a visible relationship between the longitudinal gradient from west to east and relative levels of population cluster assignment, suggesting a cline of genetic introgression of western and eastern populations. We find that as the complexity of our model increases from $K = 2$ to $K = 4$, so does evidence of increased structure in the eastern population (Fig. 3D). Additionally, the more complex model recovered a genetic cluster that is endemic to the Inter-Pecos region and northern Chihuahuan Desert (Fig. 3E), consistent with the hypothesis that the eastern population has been long-established in this area (e.g. prior to the LGM). To further test this, we examined the relative frequencies of private alleles (from nuclear SNP data) for the entire eastern population used for diversity estimates versus a subset that falls within this more restricted geographic zone of high mitochondrial diversity. Because these compared sets of individuals differ in the number of loci available after filtering and in the number of individuals, we standardized our comparison by the number of loci and numbers of individuals to obtain a relative frequency of private alleles. We find that the entire eastern population sampling (for nuclear data) has a frequency of approximately 0.17 private alleles per individual per locus, while the eastern sampling limited to the inferred endemic diversity region has a frequency of 0.20, corresponding to a 17% increase in private alleles in this region. Thus the mitochondrial and nuclear data broadly agree in identifying a region of high endemic diversity in the eastern population centered in Texas (e.g., Fig. 2B).

We found evidence for variable introgression across the genome for the admixed population of *C. atrox* in our bgc analyses (Fig. 4). In particular, we found median estimates of the genomic cline center parameter α to vary moderately among the loci sampled (minimum = - 0.0127, maximum = 0.0147) and the 95% confidence interval for α encompassed zero for all loci,

indicating a lack of excess ancestry from either parental population. The genomic cline rate parameter β was slightly less variable (minimum = - 0.0108, maximum = 0.0116) and also had a 95% confidence interval that encompassed zero for all loci. Collectively, the variation in these parameters is less extreme than that observed in other recent studies (e.g., (Gompert et al. 2012a; Lindtke et al. 2012), and would indicate that, though genome wide variation in introgression is occurring in *C. atrox*, the degree to which parental population ancestry is observed appears to be driven by drift over selection (which would be inferred in the case of large proportions of excess ancestry from one parental population over the other).

Relationship between mitochondrial and nuclear genomes

In addition to the formal tests of genomic introgression in our nuclear dataset, we examined the longitudinal relationship of our mitochondrial and nuclear genetic datasets to identify visible clines or shifts in genetic composition across their distribution. We combined samples from both datasets into two-degree longitudinal bins and calculated eastern versus western mitochondrial haplotype frequencies (per longitudinal bin), as well as average western or eastern nuclear genome proportion (inferred via population assignments of individuals under the two population model in Structure). We then plotted these values to compare nuclear and mitochondrial patterns (Fig. 5). We find a relatively steep gradient of nuclear composition shift between 110 and 108 degrees W, corresponding to the Continental Divide (and approximately the Arizona-New Mexico border), supporting this region as a barrier to western nuclear alleles moving eastward and vice versa. In contrast, the region of mitochondrial overlap is considerably broader and the shift between higher frequency western or eastern haplotypes is less steep than the shift observed in the nuclear data (Fig. 5). There is also an immediate decrease to zero in observed western mitochondrial haplotypes east of the Inter-Pecos region (east of the Pecos River), and the same

pattern is observed for eastern haplotypes west of the Continental Divide. While the greatest degree of shifting nuclear genome composition is observed within the discrete region mentioned above, there is evidence of small amounts of nuclear introgression throughout the entire longitudinal gradient. Thus, while mitochondrial introgression is confined to within the Inter-Pecos region, nuclear gene flow appears to penetrate far beyond the western and eastern geographic barriers of this region.

To test the hypothesis of unbiased gene flow between eastern and western populations of *C. atrox*, we compared mitochondrial and nuclear estimates of ancestral genetic origins of individuals for which we had both datasets (Fig. 6). If gene flow is completely unrestricted between populations we would expect to see mitochondrial haplotypes from one lineage existing with a range of different nuclear backgrounds (i.e., individuals belonging to the western mitochondrial clade may have a range of proportions of alleles from western and eastern nuclear clusters). We might also expect to see this range weighted such that nuclear genetic composition would tend to agree with the mitochondrial clade an individual belongs to, including, for example, western mitochondrial clade individuals having a higher proportion of western nuclear cluster alleles.

We find that there is a lack of high proportion western nuclear allele individuals with eastern mitochondrial haplotypes (Fig. 6). We do see evidence of the reverse, however, with several low proportion western nuclear allele individuals belonging to the western mitochondrial clade. This pattern is particularly interesting given the close geographic proximity of many of our samples, especially those within the western Inter-Pecos region (Fig. 3F). Eastern mitochondria only occur with roughly 50% or less western nuclear genome proportion, and most often possess a much lower frequency of western alleles (< 20%). Even at the steep cline of gene flow between eastern

and western populations at the Continental Divide (Fig. 3F), it is notable that eastern nuclear alleles seem to penetrate this barrier while eastern mitochondrial haplotypes do not.

Discussion

Evidence of two distinct yet introgressing lineages within C. atrox

We find evidence from both mitochondrial and nuclear datasets that *Crotalus atrox* comprises two well-differentiated lineages. Indeed, phylogenetic and network analyses of our mitochondrial dataset, and Bayesian clustering analyses of our nuclear SNP dataset collectively highlight two distinct groups of *C. atrox*, and our estimate of nuclear F_{ST} (0.15) further argues for the distinction of these lineages. We also find evidence that these lineages have introgressed following their initial isolation during the Pleistocene. The mitochondrial zone of introgression occurs across the Inter-Pecos, a broad region stretching from the Continental Divide and the Pecos River in Texas. We find explicit evidence of admixture among individuals in our nuclear SNP data, such that the majority of individuals have partial assignment to at least two population clusters under both 2 and 4 population models. Interestingly, evidence for nuclear introgression expands beyond the Inter-Pecos region on either side of the mitochondrial introgression zone, though we find the greatest degree of admixture between western and eastern clusters closest to the Continental Divide, at the western edge of the Inter-Pecos (Fig. 5).

Pleistocene isolation, expansion and a broad zone of introgression

It is evident from our sampling that individuals belonging to both eastern and western clades do co-occur, despite evidence of historical isolation. While the Continental Divide has previously been thought to be the major barrier separating eastern and western lineages, our data show that the introgression zone between populations is much larger than previously thought. Our

mitochondrial data show contrasting patterns of genetic structure in eastern and western populations, with eastern mitochondrial haplotypes tending to be more localized than western haplotypes. In contrast, we found a single western haplotype that extends from California to West Texas, consistent with recent population expansion inferred from our demographic analyses (Supplementary Online Figures 1-2). Our landscape diversity estimates show particularly low genetic diversity in both lineages near the Continental Divide, and throughout much of the Inter-Pecos region (Fig. 2A-B), consistent with evidence of population expansion in both lineages (Supplementary Online Figure 1) following Pleistocene glacial cycles.

Last Glacial Maximum (LGM) Pleistocene models suggest that eastern and western populations were separated by a wide central plateau in the Inter-Pecos region. This corroborates our inferences that the two lineages diverged in isolation for a period during the Pleistocene, prior to expanding their ranges to meet in the Inter-Pecos region since the LGM. These data together with our population genetic data indicate recent secondary contact after range expansion of both eastern and western lineages of *C. atrox* across the Inter-Pecos region. Interestingly, our LGM model predicts that the Inter-Pecos region was partially inhabited by the eastern population during the LGM. This suggests that the eastern lineage may have been present across much of the Inter-Pecos region prior to expansion of western lineages into this region, which is supported by high levels of endemic nuclear and mitochondrial diversity in some eastern portions of this region (Fig. 2B, 3D). Given that our estimate of time since population splitting greatly predates the LGM, it is possible that secondary contact between these lineages could have occurred multiple times during recessions in glaciation cycles, though this would not necessarily alter our inference of isolation followed by secondary (or tertiary) mixing.

Our results reject the hypothesis that population genetic diversity is similar between eastern and western populations of *C. atrox*. We find substantially higher levels of local genetic diversity in the eastern population relative to the western population based on nuclear SNPs (Fig. 3A-C), and mitochondrial DNA (Fig. 2A-B). It is likely that multiple factors have influenced the contrasting levels of diversity we observe, including historical demography, past and present population range size, and the diversity of habitats that occur within the distribution of each population. The eastern population is predicted to have a larger and non-linear range during the LGM, suggesting Pleistocene population sizes may have been larger in the east. Additionally, the eastern population currently occupies a region of substantially higher habitat diversity relative to the west, including forest, grasslands, and Chihuahuan Desert, in contrast to the western population that occurs primarily within arid desert habitat (Campbell and Lamar 2004).

The oldest known fossil of *C. atrox* dates from between 3.7 and 3.2 MYA (Holman 2000) and was found in north central Texas, while fossils found west of the Continental Divide are restricted to much later dates during the Late Pleistocene (Holman 1995). These fossil data suggest that the eastern population may have existed in its current range for millions of years, and substantially longer than the western population has. These data may also indicate that the range currently occupied by the eastern population is the ancestral range for the species prior to isolation of western and eastern populations due to the waxing and waning of glaciation during the Pleistocene. These data also suggest that the relative age and stability of the eastern population may have contributed to its greater genetic structure and diversity.

Evidence for incipient yet failed speciation, and sex-biased gene flow

The study of genetic incompatibilities early in the speciation continuum is essential to understanding mechanisms that drive speciation because they provide greater insight into

primary causative mechanisms that lead to reproductive isolation (Orr 1995). Given broad evidence for two divergent lineages within *C. atrox* that appear to have evolved in isolation until recent secondary contact, we were interested to test for evidence of non-random gene flow that might indicate the evolution of reproductive isolation between these lineages. While our mitochondrial and nuclear data broadly agree that two distinct lineages of *C. atrox* are currently exchanging genes across a relatively large region in the center of the species' range, patterns of introgression differ between these two datasets. While the mitochondrial introgression zone appears to be flanked by two hard boundaries – the Continental Divide to the west and the Pecos River to the east – our nuclear data indicates mixing of alleles that extends far beyond these boundaries. Nuclear data also provide evidence for a remarkably steep gradient of genotypic composition at the Continental Divide.

Assuming random mating and equal fitness of offspring, we would expect that mitochondrial genotypes might predict nuclear genotypes. For example, if an individual possesses an eastern mitochondrial haplotype it would be more likely that it contained a substantial proportion of eastern nuclear alleles, and vice-versa. However, sex-biased gene flow, as well as selection against certain combinations of mitochondrial and nuclear genotypes, may alter the expected relationships between mitochondrial and nuclear genotypes. We tested this relationship and found it to be notably asymmetrical, with western mitochondrial haplotypes associated with a wide range of western nuclear allelic content, yet eastern mitochondrial haplotypes only associated with 50% or greater eastern nuclear allelic content (Fig. 6). Thus, we did not observe individuals with eastern mitochondrial haplotypes with greater than 50% western nuclear allelic content; in most cases, eastern mitochondrial haplotypes were paired with much lower levels of western nuclear alleles. These data are consistent with a low frequency of successful eastern

female x western male mating, compared to high frequencies of western females mating with males from either population, which might be explained by sex-biased dispersal. Sex-biased gene flow is likely in snakes, given evidence for sex-biased dispersal (Keogh et al. 2007; Dubey et al. 2008; Lane and Shine 2011a). In rattlesnakes, males tend to be more widely dispersing (Duvall et al. 1992), however, a recent study found higher than expected female dispersal specifically in *C. atrox* (Schuett et al. 2013). Our results show relatively high frequencies of western haplotypes, yet low frequencies of western nuclear allelic content throughout much of the introgression zone. If this pattern were generated by sex-biased gene flow, it would require that western females have substantially higher dispersal capabilities than western males, and that dispersal for the eastern males and females is roughly equal.

A second intriguing possibility is that these results are due to a post-zygotic mechanism in which mito-nuclear incompatibilities lower the fitness of offspring that possess eastern mitochondrial haplotypes and a genetic background with high proportions of western nuclear alleles. Such an incompatibility might also be responsible for creating the steep geographic cline of allelic content at the Continental Divide (Fig. 3F), selecting against higher proportions of western alleles in populations with higher proportions of eastern mitochondria. Mito-nuclear incompatibilities have been shown in other species to be early causative drivers of post-zygotic genetic isolation and speciation (Ulloa et al. 1995; Bogdanova 2007; Presgraves 2010). It would be interesting in future studies to test competing hypotheses for what mechanisms might be driving these patterns of mito-nuclear genotypic content in *C. atrox*, to determine if there are genetic incompatibilities that limit gene flow between eastern and western *C. atrox* populations. A potential difficulty with this hypothesis is the lack of evidence for introgression driven by selection in sampled nuclear loci within the admixed *C. atrox* population. The likelihood of one

of these neutral loci being physically linked to a putative locus responsible for mito-nuclear incompatibility, however, is very low and further investigation is warranted to confirm the presence and nature of such an incompatibility. The divergence of the two main lineages of *C. atrox*, together with evidence for biased gene flow and potential post-zygotic isolating mechanisms, argue that these two lineages were at some intermediate stages along the speciation continuum prior to their secondary extensive mixing.

Based on our findings, we conclude that *C. atrox* represents a single species, comprised of two divergent populations that are experiencing ongoing widespread gene flow that extends essentially to the margins of their entire distribution. Our results support the hypothesis that, at some time in the past, *C. atrox* comprised two well-differentiated lineages that represented incipient species that were each evolving in partial genetic isolation. While this extended period of mutual isolation placed these two lineages at an intermediate stage of the speciation continuum, subsequent pervasive gene flow, presumably after substantial geographic population expansion following the Pleistocene, has apparently reversed the progression of these two lineages along this continuum, and broadly mixed genetic variation between these lineages. The extent of gene flow among lineages we have observed in this study leads to the consideration of *C. atrox* as a single species based on a number of species concepts, including the biological (Mayr 1963), the evolutionary (Simpson 1951; Wiley 1978), and the general lineage concept (de Queiroz 1998). Recently, species concepts that recognize speciation with gene flow have been favored in some cases (Feder et al. 2012), and there is empirical evidence that such processes may be fairly common in nature (Nosil 2008; Pinho and Hey 2010). Speciation with gene flow has, however, been primarily documented in instances where gene flow is confined to localized regions of sympatry and peripatry (Leache et al. 2013; Martin et al. 2013; Osborne et al. 2013).

The broad extent and penetrance of gene flow across nearly the entire range of C. atrox, together with evidence that these two lineages appear to be expanding and, if anything, mixing to a greater extent as time progresses, suggests that *C. atrox* represents an example of failed speciation rather than an example of speciation with gene flow.

Acknowledgments

We thank Jens Vindum and the California Academy of Sciences, Robert Murphy and the Royal Ontario Museum, Jesse Meik, and Corey Roelke for providing tissue samples; Jill Castoe, Rachel Wostl, and Nicole Hales for help with laboratory work; and Matthew Fujita for helpful Perl scripts. We thank Jeff Streicher and Eric Watson for comments and suggestions. Support for this work was provided by faculty startup funds from the University of Texas at Arlington to TAC and by a Phi Sigma Beta Phi Chapter research grant to DRS.

Figures

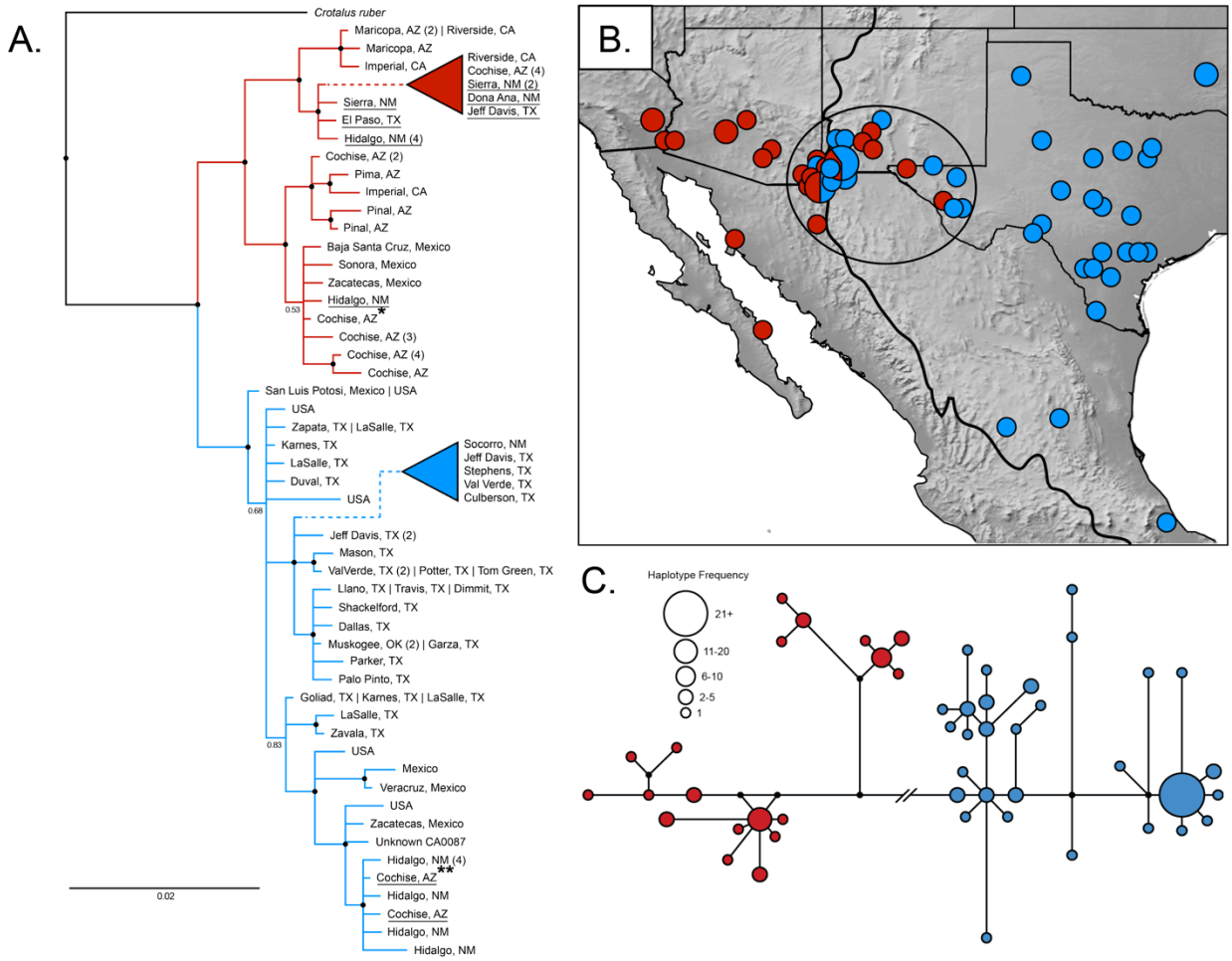


Figure 1. Results of analyses of the ND4 mitochondrial gene. **A.** Bayesian phylogenetic tree estimate of relationships among condensed *C. atrox* haplotypes. The western clade is represented by red branches and the eastern clade by blue branches. Bipartition posterior probabilities greater than 0.90 are represented by black circles. Haplotypes marked with * and ** include 16 and 49 individuals, respectively. Samples falling within the western or eastern clade counter to geographic expectation are underlined. **B.** Map of western (red circles) and eastern (blue circles) clades, and the Continental Divide (dark line). The size of each circle corresponds to sampling frequency at that locality. For localities with pie charts, the relative contribution of each color to the circle reflects the frequency of eastern or western clade individuals. The black ellipse highlights the putative fusion region between eastern and western clades. **C.** Median-joining network of mtDNA haplotypes. The two mitochondrial clades are represented by blue (eastern) and red (western) circles. Circle sizes correspond to the frequency of each haplotype in our sampling. Branch lengths between circles are proportional to the number of nucleotide differences between adjacent haplotypes. Black circles represent a haplotype that is absent in our sampling.

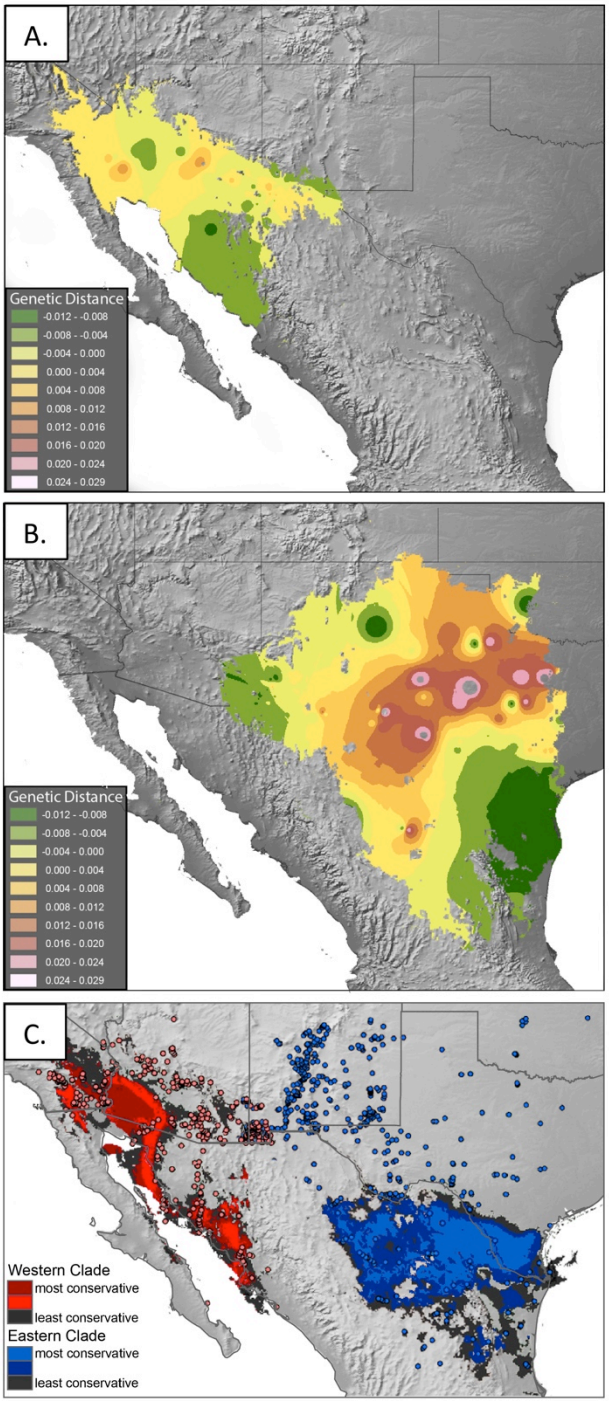


Figure 2. Landscape genetic diversity estimates and Last Glacial Maximum (LGM) ecological niche models. **A-B.** Residual pairwise genetic distances from the calculated linear regression interpolated across landscape for western (**A**) and eastern (**B**) mitochondrial clades of *C. atrox*. Interpolations are restricted by ecological niche models used to determine suitable habitat for each population and by the distribution of sampled localities. **C.** LGM models for western (red) and eastern (blue) populations at three stringency thresholds.

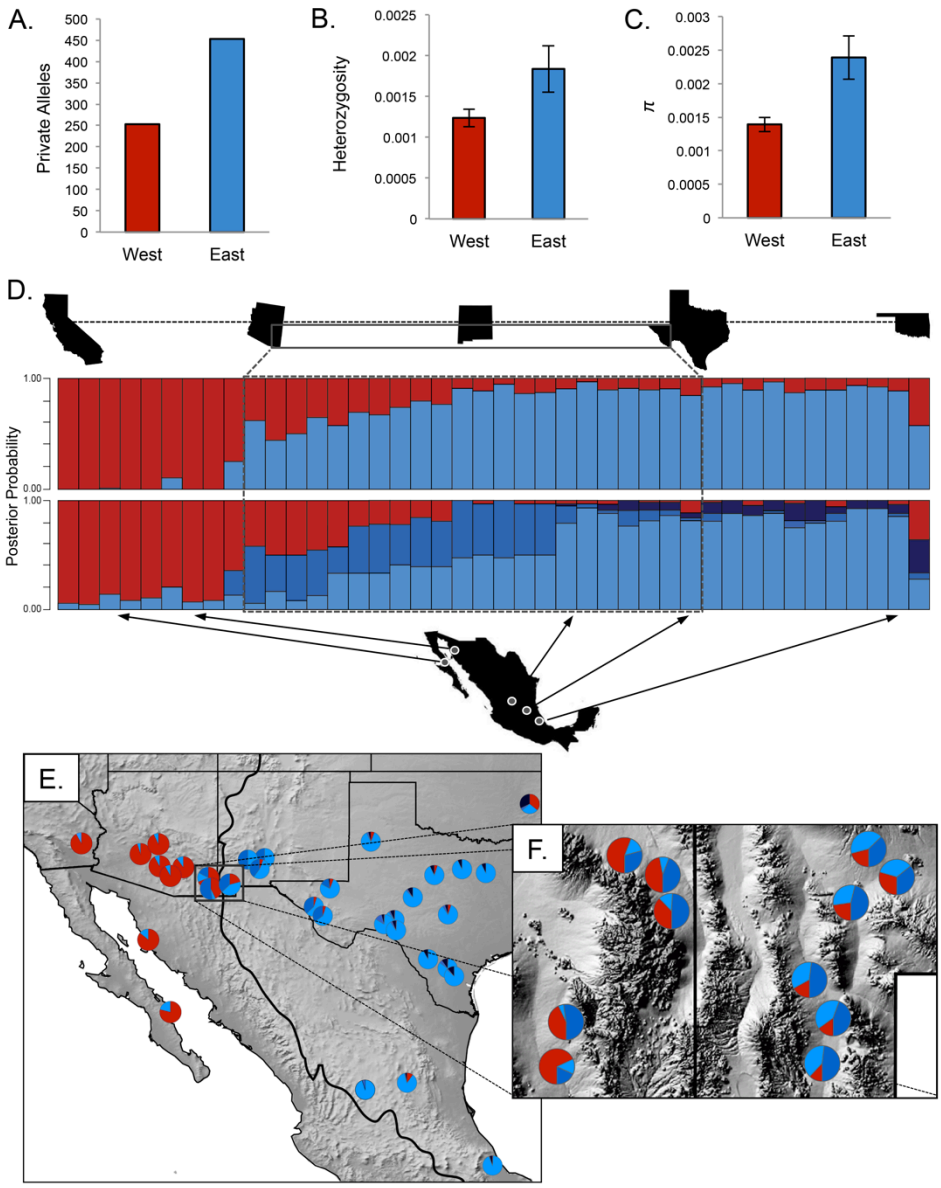


Figure 3. Patterns of genetic structure and diversity based on nuclear SNP data. **A., B., and C.** Estimates of private alleles, heterozygosity, and nucleotide diversity (respectively) for western and eastern populations inferred from analyses in *Stacks*. Error bars represent the variance in estimates of heterozygosity and nucleotide diversity. **D.** Structure plots for K=2 (top) and K=4 (bottom) population cluster models. Individuals are represented by individual bars of the posterior probability of their assignment to various population clusters. Individuals are ordered longitudinally from west to east. The putative region of lineage fusion (introgression) is highlighted in grey, and the dashed grey box includes samples that fall within this region. **E.** Map of samples used in nuclear SNP analyses. Individuals are represented by pie charts with colors corresponding to the structure plot for the K = 4 model, and the Continental Divide is represented by the bold black line. The grey box highlights a region of dense sampling at the continental divide, which is shown in more detail in the inset to the right (**F**).

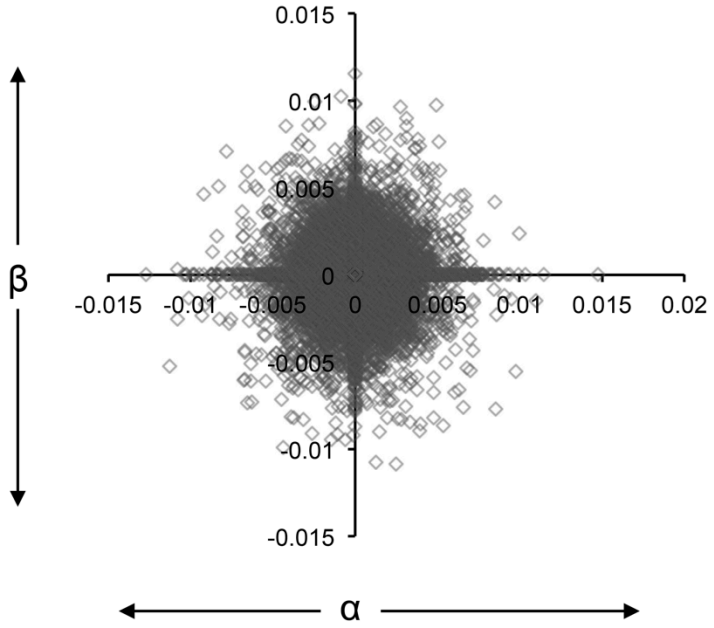


Figure 4. Variable levels of introgression in the admixed *C. atrox* population as determined by the genomic cline center parameter α and the genomic cline rate parameter β estimated in bgc for each nuclear locus. Each individual locus is represented by a grey box.

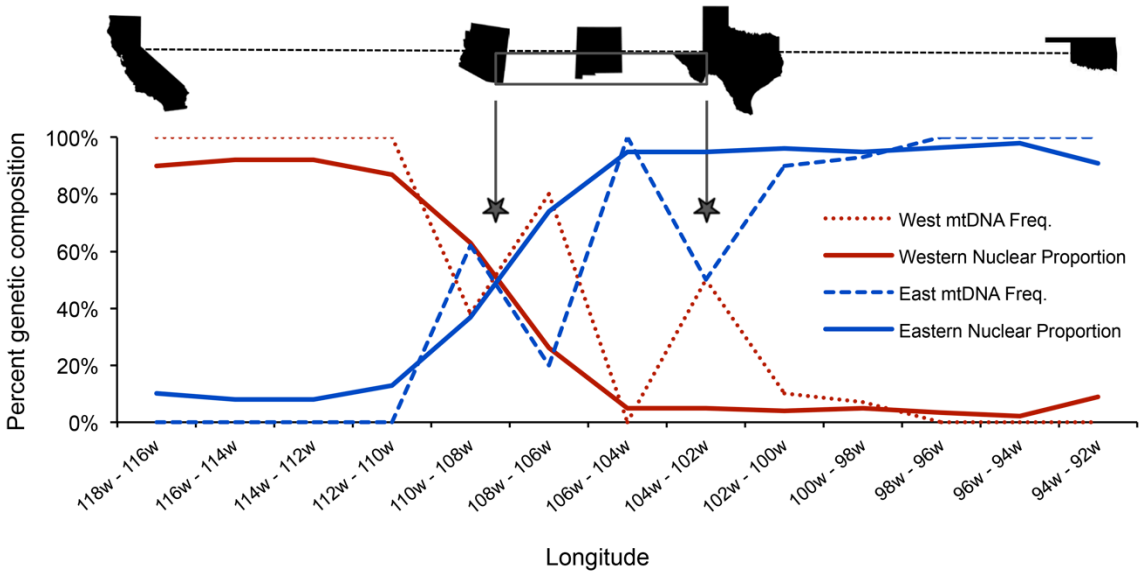


Figure 5. Clines of mitochondrial and nuclear genetic composition across a longitudinal gradient. Estimates of genetic composition are organized in two-longitudinal-degree bins from west to east. Values of western and eastern nuclear allele proportions per longitudinal bin are represented by red and blue solid lines, respectively. Western and eastern mitochondrial haplotype frequencies per longitudinal bin are represented by dashed red and blue lines. Points of interest (the Continental Divide and eastern extreme of the Inter-Pecos region) are highlighted with black stars, which also highlight notable shifts in nuclear or mitochondrial genetic content.

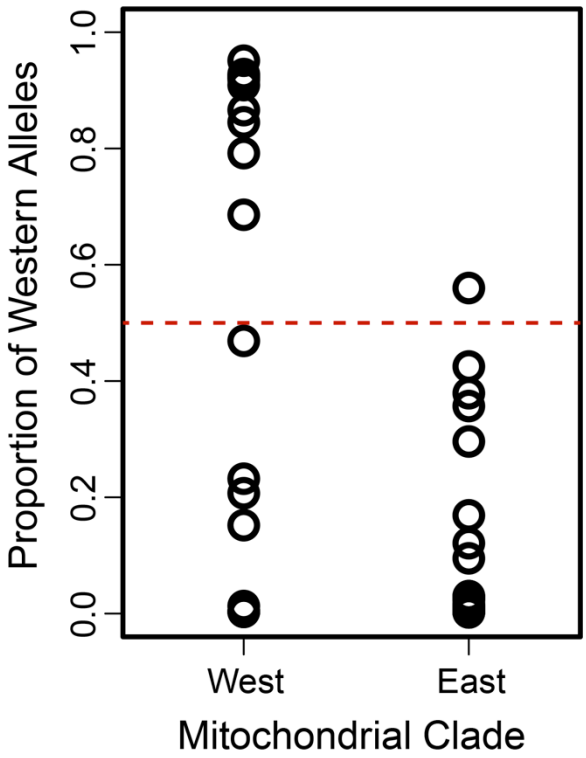
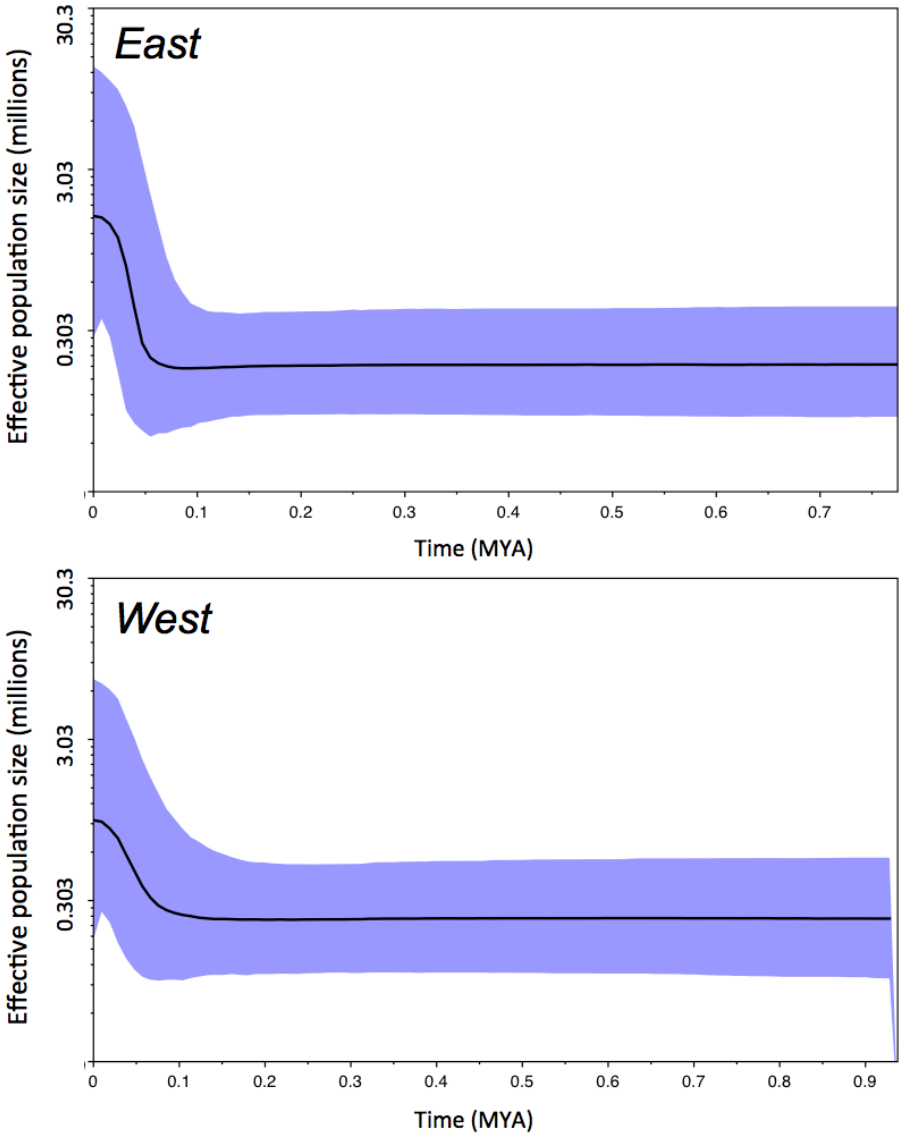
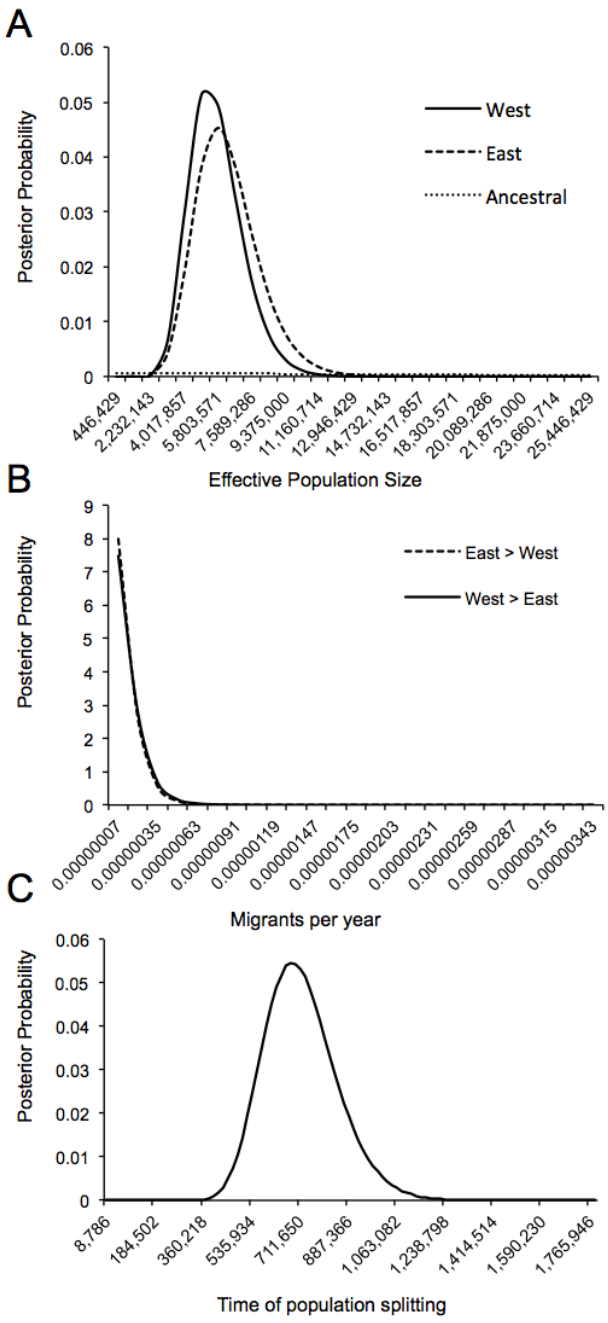


Figure 6. Deficiency of individuals with eastern mitochondrial haplotypes and moderate to high levels of western nuclear alleles. Comparison of mitochondrial haplotype clade membership with relative proportion of western nuclear alleles inferred from *Structure* analysis based on a two population model. The red line indicates a 50% threshold of western alleles



Supplementary Figure 1. Bayesian skyline plots based on mitochondrial data from the east and west clades of *Crotalus atrox*, which estimate historical dynamics in effective population size (in millions) for each lineage, shown on a log scale. Confidence intervals for estimates of effective population size through time are represented by the violet background surrounding the solid black line representing the mean estimate. Demographic units of population size were calculated using a generation time estimate of 3.3 years.



Supplementary Figure 2. Results of isolation-with-migration model parameter estimation in IMa2. **A.** Marginal posterior probability density curves for effective population sizes of the eastern, western, and ancestral populations. **B.** Migration rate (in migrants per year) curves for eastern and western populations. **C.** Time since splitting of the ancestral population into eastern and western populations. All parameter distributions were converted to demographic units using mutation rate for the ND4 gene and generation time estimates from Castoe et al. (2007).

Supplementary Table 1. Voucher, locality, and GenBank accession information for specimens used in this study. Data (i.e., ND4 or both ND4 and RADseq) used for each specimen is listed. Where ND4 is marked with a * or **, the specimen belongs to that respective haplotype group on the mitochondrial gene tree (Fig. 1A).

Species	CA Number	Voucher	Country	State	County	Data	GenBank	RAD Accessi	Reference
Crotalus molossus	-	CLP66	USA	TX	El Paso	ND4	AY223695		Castoe et al. 2007
Crotalus tigris	-	CLP169	USA	AZ	Pima	ND4	AF156574		Castoe et al. 2007
Crotalus ruber	-	RWV2001-08	USA	CA	RIverside	ND4	DQ679838		Castoe et al. 2007
Crotalus atrox	CA0083	ROM18244	Mexico	Baja California	-	ND4 + RADs	DQ679840	SRR1706413	Castoe et al. 2007
	CA0031	ENS10537	Mexico	San Luis Potosi	-	ND4 + RADs	DQ679842	SRR1706414	Castoe et al. 2007
	CA0079*	LSUMZ H-5548	Mexico	Sonora	-	ND4 + RADs	DQ679844	SRR1706415	Castoe et al. 2007
	CA0075	TWR 1249	Mexico	Sonora	-	ND4	DQ679843		Castoe et al. 2007
	CA0032	ENS10538	Mexico	Veracruz	-	ND4 + RADs	DQ679845	SRR1706416	Castoe et al. 2007
	CA0033	ENS10539	Mexico	Zacatecas	-	ND4 + RADs	DQ679846	SRR1706417	Castoe et al. 2007
	CA0034	ENS10540	Mexico	Zacatecas	-	ND4	KP250883		This study
	CA0035	ENS10536	Mexico	San Luis Potosi	-	ND4	DQ679841		Castoe et al. 2007
	CA0122	CLS210	USA	AZ	Cochise	ND4	KP250884		This study
	CA0191	CLS368	USA	AZ	Cochise	ND4	KP250885		This study
	CA0165*	CLS305	USA	AZ	Cochise	ND4	KP250886		This study
	CA0155*	CLS293	USA	AZ	Cochise	ND4	KP250887		This study
	CA0168*	CLS308	USA	AZ	Cochise	ND4	KP250888		This study
	CA0156**	CLS294	USA	AZ	Cochise	ND4	KP250889		This study
	CA0185	CLS388	USA	AZ	Cochise	ND4	KP250890		This study
	CA0119*	CLS418	USA	AZ	Cochise	ND4	KP250891		This study
	CA0124	CLS212	USA	AZ	Cochise	ND4	KP250892		This study
	CA0112	CLS384	USA	AZ	Cochise	ND4	KP250893		This study
	CA0164	CLS304	USA	AZ	Cochise	ND4	KP250894		This study
	CA0179	CLS346	USA	AZ	Cochise	ND4	KP250895		This study
	CA0157	CLS297	USA	AZ	Cochise	ND4	KP250896		This study
	CA0118*	CLS416	USA	AZ	Cochise	ND4	KP250897		This study
	CA0171*	CLS344	USA	AZ	Cochise	ND4 + RADs	KP250898	SRR1706418	This study
	CA0123	CLS211	USA	AZ	Cochise	ND4	KP250899		This study
	CA0116**	CLS413	USA	AZ	Cochise	ND4 + RADs	KP250900	SRR1706419	This study
	CA0177**	CLS341	USA	AZ	Cochise	ND4	KP250901		This study
	CA0178*	CLS342	USA	AZ	Cochise	ND4	KP250902		This study
	CA0121*	CLS425	USA	AZ	Cochise	ND4	KP250903		This study
	CA0181**	CLS347	USA	AZ	Cochise	ND4	KP250904		This study
	CA0180*	CLS348	USA	AZ	Cochise	ND4	KP250905		This study
	CA0174**	CLS329	USA	AZ	Cochise	ND4	KP250906		This study
	CA0115*	CLS404	USA	AZ	Cochise	ND4	KP250907		This study
	CA0172**	CLS328	USA	AZ	Cochise	ND4	KP250908		This study
	CA0175**	CLS330	USA	AZ	Cochise	ND4	KP250909		This study
	CA0117**	CLS414	USA	AZ	Cochise	ND4	KP250910		This study
	CA0125**	CLS217	USA	AZ	Cochise	ND4	KP250911		This study
	CA0138**	CLS260	USA	AZ	Cochise	ND4	KP250912		This study
	CA0151	CLS290	USA	AZ	Cochise	ND4 + RADs	KP250913	SRR1706420	This study
	CA0127**	CLS226	USA	AZ	Cochise	ND4	KP250914		This study
	CA0128	CLS233	USA	AZ	Cochise	ND4	KP250915		This study
	CA0146**	CLS282	USA	AZ	Cochise	ND4 + RADs	KP250916	SRR1706421	This study
	CA0169**	CLS309	USA	AZ	Cochise	ND4	KP250917		This study
	CA0182**	CLS352	USA	AZ	Cochise	ND4	KP250918		This study
	CA0140**	CLS265	USA	AZ	Cochise	ND4	KP250919		This study
	CA0192	CLS369	USA	AZ	Cochise	ND4	KP250920		This study
	CA0120**	CLS419	USA	AZ	Cochise	ND4	KP250921		This study
	CA0126**	CLS218	USA	AZ	Cochise	ND4	KP250922		This study
	CA0142**	CLS274	USA	AZ	Cochise	ND4	KP250923		This study
	CA0137**	CLS259	USA	AZ	Cochise	ND4	KP250924		This study
	CA0152**	CLS291	USA	AZ	Cochise	ND4	KP250925		This study
	CA0170**	CLS315	USA	AZ	Cochise	ND4	KP250926		This study
	CA0139**	CLS264	USA	AZ	Cochise	ND4 + RADs	KP250927	SRR1706422	This study
	CA0038	CWP2	USA	AZ	Cochise	ND4	DQ679847		Castoe et al. 2007
	CA0144	CLS279	USA	AZ	Cochise	ND4	KP250928		This study
	CA0145	CLS280	USA	AZ	Cochise	ND4	KP250929		This study
	CA0129	CLS235	USA	AZ	Cochise	ND4	KP250930		This study
	CA0009*	CLS494	USA	AZ	Cochise	ND4	DQ679849		Castoe et al. 2007
	CA0006*	CLS474	USA	AZ	Cochise	ND4	DQ679848		Castoe et al. 2007
	CA0044	ENT2	USA	AZ	Maricopa	ND4	DQ679850		Castoe et al. 2007
	CA0046	ENT11	USA	AZ	Maricopa	ND4 + RADs	DQ679851	SRR1706423	Castoe et al. 2007

CA0048	ENT7	USA	AZ	Maricopa	ND4 + RADs	DQ679852	SRR1706424	Castoe et al. 2007
CA0049	LACM150957	USA	AZ	Pima	ND4 + RADs	DQ679853	SRR1706425	Castoe et al. 2007
CA0042	ENTA21	USA	AZ	Pinal	ND4 + RADs	DQ679854	SRR1706426	Castoe et al. 2007
CA0043	ENTA49	USA	AZ	Pinal	ND4 + RADs	DQ679855	SRR1706427	Castoe et al. 2007
CA0013	RNF2596	USA	CA	Imperial	ND4	DQ679857		Castoe et al. 2007
CA0012	RNF2595	USA	CA	Imperial	ND4	DQ679856		Castoe et al. 2007
CA0081	ROM2398	USA	CA	Riverside	ND4	DQ679858		Castoe et al. 2007
CA0082	ROM18144	USA	CA	Riverside	ND4 + RADs	DQ679859	SRR1706428	Castoe et al. 2007
CA0007**	CLS482	USA	NM	Grant	ND4	DQ679861		Castoe et al. 2007
CA0107**	CLS379	USA	NM	Hidalgo	ND4	KP250931		This study
CA0166**	CLS306	USA	NM	Hidalgo	ND4	KP250932		This study
CA0130*	CLS240	USA	NM	Hidalgo	ND4 + RADs	KP250933	SRR1706429	This study
CA0105	CLS377	USA	NM	Hidalgo	ND4	KP250934		This study
CA0106**	CLS378	USA	NM	Hidalgo	ND4	KP250935		This study
CA0131**	CLS243	USA	NM	Hidalgo	ND4	KP250936		This study
CA0114**	CLS393	USA	NM	Hidalgo	ND4	KP250937		This study
CA0143	CLS277	USA	NM	Hidalgo	ND4	KP250938		This study
CA0150	CLS287	USA	NM	Hidalgo	ND4	KP250939		This study
CA0135	CLS250	USA	NM	Hidalgo	ND4	KP250940		This study
CA0136**	CLS254	USA	NM	Hidalgo	ND4	KP250941		This study
CA0102**	CLS373	USA	NM	Hidalgo	ND4	KP250942		This study
CA0132	CLS244	USA	NM	Hidalgo	ND4 + RADs	KP250943	SRR1706430	This study
CA0111	CLS383	USA	NM	Hidalgo	ND4 + RADs	KP250944	SRR1706431	This study
CA0189**	CLS364	USA	NM	Hidalgo	ND4	KP250945		This study
CA0187**	CLS362	USA	NM	Hidalgo	ND4	KP250946		This study
CA0104**	CLS376	USA	NM	Hidalgo	ND4	KP250947		This study
CA0133**	CLS247	USA	NM	Hidalgo	ND4	KP250948		This study
CA0188	CLS363	USA	NM	Hidalgo	ND4	KP250949		This study
CA0103**	CLS374	USA	NM	Hidalgo	ND4	KP250950		This study
CA0005**	CLS471	USA	NM	Hidalgo	ND4	DQ679863		Castoe et al. 2007
CA0183	CLS354	USA	NM	Hidalgo	ND4	KP250951		This study
CA0162**	CLS302	USA	NM	Hidalgo	ND4	KP250952		This study
CA0141**	CLS268	USA	NM	Hidalgo	ND4	KP250953		This study
CA0148**	CLS284	USA	NM	Hidalgo	ND4	KP250954		This study
CA0158**	CLS298	USA	NM	Hidalgo	ND4 + RADs	KP250955	SRR1706432	This study
CA0159**	CLS299	USA	NM	Hidalgo	ND4	KP250956		This study
CA0110	CLS382	USA	NM	Hidalgo	ND4 + RADs	KP250957	SRR1706433	This study
CA0149	CLS286	USA	NM	Hidalgo	ND4	KP250958		This study
CA0160	CLS300	USA	NM	Hidalgo	ND4 + RADs	KP250959	SRR1706434	This study
CA0161**	CLS301	USA	NM	Hidalgo	ND4	KP250960		This study
CA0163	CLS303	USA	NM	Hidalgo	ND4	KP250961		This study
CA0190**	CLS366	USA	NM	Hidalgo	ND4	KP250962		This study
CA0108**	CLS380	USA	NM	Hidalgo	ND4	KP250963		This study
CA0109**	CLS381	USA	NM	Hidalgo	ND4	KP250964		This study
CA0134**	CLS249	USA	NM	Hidalgo	ND4	KP250965		This study
CA0047	ENT12	USA	NM	Hidalgo	ND4	DQ679862		Castoe et al. 2007
CA0022	RWV2001-14	USA	NM	Dona Ana	ND4 + RADs	DQ679864	SRR1706435	Castoe et al. 2007
CA0021	RWV2001-13	USA	NM	Sierra	ND4 + RADs	DQ679860	SRR1706436	Castoe et al. 2007
CA0039	BLC27	USA	NM	Sierra	ND4 + RADs	DQ679865	SRR1706437	Castoe et al. 2007
CA0041	BLC	USA	NM	Sierra	ND4	DQ679866		Castoe et al. 2007
CA0040	BLC	USA	NM	Socorro	ND4	DQ679867		Castoe et al. 2007
CA0336	LNV336	USA	OK	Muskogee	ND4 + RADs	KP250966	SRR1706438	This study
CA0337	LNV337	USA	OK	Muskogee	ND4	KP250967		This study
CA0099	RLG390	USA	TX	Dimmit	ND4 + RADs	KP250968	SRR1706439	This study
CA0057	TJL566	USA	TX	Duval	ND4	DQ679870		Castoe et al. 2007
CA0058	TJL588	USA	TX	Goliad	ND4	DQ679873		Castoe et al. 2007
CA0197	JWS656	USA	TX	Jeff Davis	ND4	KP250969		This study
CA0052	CLP64	USA	TX	Jeff Davis	ND4	DQ679876		Castoe et al. 2007
CA0061	TJL719	USA	TX	Karnes	ND4	DQ679878		Castoe et al. 2007
CA0098	RLG380	USA	TX	LaSalle	ND4 + RADs	KP250970	SRR1706440	This study
CA0096	RLG381	USA	TX	LaSalle	ND4 + RADs	KP250971	SRR1706441	This study
CA0020	RWV2001-12	USA	TX	LaSalle	ND4	DQ679879		Castoe et al. 2007
CA0194	DRS0003	USA	TX	Palo Pinto	ND4 + RADs	KP250972	SRR1706442	This study
CA0195	DRS0005	USA	TX	Parker	ND4 + RADs	KP250973	SRR1706443	This study
CA0193	DRS0002	USA	TX	Shackelford	ND4 + RADs	KP250974	SRR1706444	This study
CA0196	DRS0007	USA	TX	Tom Green	ND4 + RADs	KP250975	SRR1706445	This study
CA0097	RLG367	USA	TX	Val Verde	ND4 + RADs	KP250976	SRR1706446	This study
CA0100	RLG404	USA	TX	Zavala	ND4 + RADs	KP250977	SRR1706447	This study
CA0073	JJ	USA	TX	Culberson	ND4 + RADs	DQ679868	SRR1706448	Castoe et al. 2007
CA0072	JJ	USA	TX	Dallas	ND4	DQ679869		Castoe et al. 2007

CA0051	CLP60	USA	TX	El Paso	ND4	DQ679871		Castoe et al. 2007
CA0074	CLS576	USA	TX	Garza	ND4	DQ679872		Castoe et al. 2007
CA0028	RWV2001-22	USA	TX	Jeff Davis	ND4 + RADs	DQ679875	SRR1706449	Castoe et al. 2007
CA0018	RWV2001-09	USA	TX	Jeff Davis	ND4 + RADs	DQ679874	SRR1706450	Castoe et al. 2007
CA0060	TJL593	USA	TX	Karnes	ND4	DQ679877		Castoe et al. 2007
CA0062	TJL527	USA	TX	LaSalle	ND4	DQ679880		Castoe et al. 2007
CA0063	TJL601	USA	TX	Llano	ND4 + RADs	DQ679881	SRR1706451	Castoe et al. 2007
CA0064	TJL	USA	TX	Mason	ND4	DQ679882		Castoe et al. 2007
CA0065	TJL868	USA	TX	Potter	ND4 + RADs	DQ679883	SRR1706452	Castoe et al. 2007
CA0053	CLP199	USA	TX	Stephens	ND4	DQ679884		Castoe et al. 2007
CA0066	TJL718	USA	TX	Travis	ND4	DQ679885		Castoe et al. 2007
CA0069	TJL348	USA	TX	Val Verde	ND4 + RADs	DQ679887	SRR1706453	Castoe et al. 2007
CA0068	TJL347	USA	TX	Val Verde	ND4 + RADs	DQ679886	SRR1706454	Castoe et al. 2007
CA0071	TJL775	USA	TX	Zapata	ND4	DQ679888		Castoe et al. 2007
CA0085	JAC26689	USA	-	-	ND4	KP250978		This study
CA0086	JAC29236	USA	-	-	ND4	KP250979		This study
CA0087	JAC29282	USA	-	-	ND4	KP250980		This study
CA0088	JAC29568	USA	-	-	ND4	KP250981		This study
CA0089	JAC29818	USA	-	-	ND4	KP250982		This study
CA0091	JAC29854	USA	-	-	ND4	KP250983		This study
CA0147**	CLS283	USA	-	-	ND4	KP250984		This study
CA0184**	CLS385	USA	-	-	ND4	KP250985		This study

Supplementary Table 2. Genetic diversity estimates calculated from nucleotide polymorphisms (SNPs) using ddRAD sequencing of *Crotalus atrox* for our two-population model. Mean and variance for the number of variable loci, number of private alleles, heterozygosity, genetic diversity (Pi), and F_{ST} values are shown. Coverage refers to the minimum read requirement per stack coverage (5×, 10×, 20×), while threshold represents the proportion of individuals that must be represented by each locus (50% and 75%).

Coverage: 5×	**Threshold**	**Polymorphic Loci**		**Private Alleles**		**Heterozygosity**		**Pi (π)**		F_{ST}
		West	East	West	East	West	East	West	East	
Mean	50%	1438	738	253	453	0.0012	0.0018	0.0014	0.0024	0.15
	75%	718	568	117	351	0.0011	0.0018	0.0013	0.0024	0.1
Variance	50%	-	-	-	-	0.00011	0.00003	0.00012	0.00032	-
	75%	-	-	-	-	0.00014	0.00032	0.00014	0.00039	-

Coverage: 10×	**Threshold**	**Polymorphic Loci**		**Private Alleles**		**Heterozygosity**		**Pi (π)**		F_{ST}
		West	East	West	East	West	East	West	East	
Mean	50%	848	903	189	513	0.0014	0.0021	0.0014	0.0026	0.11
	75%	400	252	44	170	0.0012	0.0019	0.0012	0.0023	0.11
Variance	50%	-	-	-	-	0.00016	0.0003	0.00015	0.00034	-
	75%	-	-	-	-	0.0002	0.00051	0.00018	0.00058	-

Coverage: 20×	**Threshold**	**Polymorphic Loci**		**Private Alleles**		**Heterozygosity**		**Pi (π)**		F_{ST}
		West	East	West	East	West	East	West	East	
Mean	50%	400	335	69	213	0.0012	0.0021	0.0013	0.0025	0.12
	75%	199	63	13	49	0.0011	0.0017	0.0012	0.0021	0.09
Variance	50%	-	-	-	-	0.0002	0.0005	0.00019	0.00054	-
	75%	-	-	-	-	0.00025	0.00095	0.00025	0.0011	-

Chapter 3

Insight into the roles of selection in speciation from genomic patterns of divergence and introgression in secondary contact in venomous rattlesnakes

Drew R. Schield[1], Richard H. Adams[1], Daren C. Card[1], Blair W. Perry[1], Giulia M. Pasquesi[1], Tereza Jezkova[2], Daniel M. Portik[1], Audra L. Andrew[1], Carol L. Spencer[3], Elda E. Sanchez[4], Matthew K. Fujita[1], Stephen P. Mackessy[5], and Todd A. Castoe[1, §]

[1]Department of Biology, 501 S. Nedderman Dr., The University of Texas at Arlington, Arlington, TX 76010, USA

[2]Department of Ecology and Evolutionary Biology, 1041 E. Lowell St., University of Arizona, Tucson, AZ 85721 USA

[3]Museum of Vertebrate Zoology, 3101 Valley Life Sciences Building, University of California, Berkeley, CA 94720 USA

[4]National Natural Toxins Research Center and Department of Chemistry, 975 W. Ave. B., Texas A&M University Kingsville, Kingsville, TX 78363

[5]School of Biological Sciences, 501 20th St., University of Northern Colorado, Greeley, CO 80639, USA

Abstract

Investigating secondary contact of historically isolated lineages can provide insight into how selection and drift influence genomic divergence and admixture. Here we studied the genomic landscape of divergence and introgression following secondary contact between lineages of the Western Diamondback Rattlesnake (*Crotalus atrox*) to determine if genomic regions under selection in allopatry also contribute to reproductive isolation during introgression. We used thousands of nuclear loci to study genomic differentiation between two lineages that have experienced recent secondary contact following isolation, and incorporated sampling from a zone of secondary contact to identify loci that are resistant to gene flow in hybrids. Comparisons of patterns of divergence and introgression revealed a positive relationship between allelic differentiation and resistance to introgression across the genome, and greater than expected overlap between genes linked to lineage-specific divergence and loci that resist introgression. Genes linked to putatively selected markers were related to prominent aspects of rattlesnake biology that differ between populations of Western Diamondback rattlesnakes (i.e., venom and reproductive phenotypes). We also found evidence for selection against introgression of genes that may contribute to cytonuclear incompatibility, consistent with previously observed biased patterns of nuclear and mitochondrial alleles suggestive of partial reproductive isolation due to cytonuclear incompatibilities. Our results provide a genome-scale perspective on the relationships between divergence and introgression in secondary contact that is relevant for understanding the roles of selection in maintaining partial isolation of lineages, causing admixing lineages to not completely homogenize.

Introduction

Understanding the process of speciation requires insight into both the processes that underlie lineage divergence in isolation and the processes that maintain lineage integrity (i.e., limit gene flow) in the face of gene flow during secondary contact. Evolutionary divergence involves both selection-driven and neutral changes as lineages evolve (Coyne and Orr 2004; Kuehne et al. 2007; Nosil et al. 2009). Loci that have accumulated divergence neutrally in allopatry may eventually act to limit gene flow between sister lineages due to a reduction in hybrid fitness when introgression occurs (e.g., cytonuclear incompatibilities; (Orr 1995; Ulloa et al. 1995; Fishman and Willis 2006; Sambatti et al. 2008; Gagnaire et al. 2012). Conversely, locally adapted genes subject to geographically variable selection may also contribute to reduced gene flow and genetic isolation in structured populations that come into secondary contact (Orr and Smith 1998; Boughman 2001; Rosenblum 2006; Nosil et al. 2008), as alleles from one parental population may offer a greater fitness advantage to hybrids. Accordingly, insight into the roles of selective and neutral evolutionary processes in driving speciation can be gained from the relationships (or lack there of) between genes important in divergence and introgression in natural systems.

While links between adaptive evolution and speciation have been established (Schluter and Conte 2009; Faria et al. 2014), a largely unanswered question is whether the same genomic regions are important in both the process of divergence in isolation (i.e., loci under divergent selection and adaptation) and in preventing gene flow in secondary contact (i.e., loci underlying partial or complete reproductive isolation). This is an important question central to our understanding of how speciation proceeds (or is reversed), as there is a wealth of examples of

secondary contact and hybridization in diverse taxa (Payseur and Rieseberg 2016). Despite this, there are few studies that have specifically examined adaptive evolution in both divergence and in secondary contact (e.g., (Gompert et al. 2012a; Nosil et al. 2012; Parchman et al. 2013). These studies have explored connections between divergent selection during isolation with selection against introgression upon secondary contact, and provide evidence that loci evolved under divergent selection also contribute to partial or complete reproductive isolation during hybridization due to deleterious fitness effects on hybrids. However, these studies are limited to a small representation of taxa (two insect systems and one bird system), and it is unknown how broadly the genomic relationships between divergence and admixture processes are found in nature. There is thus a need to test the consistency of these relationships broadly across taxa to appreciate their evolutionary significance, and how such relationships may be dependent on particular taxa or on the degree of differentiation between diverged lineages.

In this paper, we address this need by examining a previously described a hybrid zone between two divergent and historically isolated lineages of the Western Diamondback Rattlesnake (*Crotalus atrox*) (Castoe et al. 2007; Schield et al. 2015). *C. atrox* inhabits a broad distribution including regions of the Chihuahuan and Sonoran deserts, as well as the adjacent temperate grasslands in the southwestern United States and northern Mexico (Campbell and Lamar 2004). The range of *C. atrox* is bisected by the Continental Divide (Fig. 1A), which represents a major constriction of two large tracts of occupied habitat to the east and west. The two distinct lineages (i.e., parental populations) of *C. atrox* generally coincide with populations east and west of the Continental Divide and the admixture region between these lineages is bounded by the Continental Divide to the west and the Pecos River to the east; this admixture region is generally dominated by Chihuahuan Desert habitat (Castoe et al. 2007; Schield et al. 2015), and we here

refer to this area as the Chihuahuan region. Additionally, ecological niche models projected onto the climatic conditions of the Last Glacial Maximum suggest that ancestral *C. atrox* populations diverged in prolonged isolation and have since converged within the admixture region that exists geographically between their two ancestral ranges (Schield et al. 2015). Using a comparison of mtDNA and nuclear SNPs, we identified an asymmetry in the proportions of nuclear alleles paired with mitochondrial haplotypes in the Chihuahuan region. Specifically, we did not observe individuals with largely 'western' nuclear genomes with 'eastern' mitochondrial haplotypes (Schield et al. 2015). We proposed that this asymmetry could have arisen from cytonuclear incompatibilities that evolved during divergence in isolation and that reduce the fitness of hybrids in secondary contact, which we explore further in this paper with expanded data and analyses.

In addition to studying the possibility of cytonuclear incompatibility between historically isolated lineages of *C. atrox,* in this paper we consider aspects of their biology that may have been driven by adaptive evolution and lineage-specific patterns of selection. Previous studies have explored phenotypic and natural history diversity across the range of *C. atrox* and, in particular, differences between eastern and western populations. For example, there are differences in body coloration and patterning between populations (Klauber 1930; Spencer 2003; Spencer 2008), with darker overall coloration and fewer body blotches found in the eastern population. Spencer (2003) also discovered differences in reproduction between eastern and western females; reproductive differences included a significant correlation between longitude and the number of follicles per female, with larger and more numerous follicles in western females, along with differences in offspring size. Finally, venom toxicity and composition is

known to vary between eastern and western populations, including increased protease activity in the eastern population (Minton and Weinstein 1986).

Our broad aim in this study was to identify genome-wide patterns of selection in divergent lineages of *C. atrox*, and test if these or different genomic loci were also under selection in zones of admixture where these lineages meet in secondary contact. To accomplish this, we first use our large nuclear locus dataset to test competing models of speciation in Western Diamondbacks to test the hypothesis of population divergence followed by secondary contact, which is strongly supported by previous studies (Castoe et al. 2007; Schield et al. 2015). We then leverage this system to address the following questions central to the role of adaptive evolution in lineage divergence and introgression in the process of secondary contact and hybridization: i) Is there evidence for loci linked to genomic regions under divergent selection for local adaptation within parental lineages/populations? And ii) Is there evidence for loci under selection to resist introgression in the hybrid population? We then test the hypothesis that there is a genomic relationship (i.e., a positive correlation) between genetic differentiation in parental populations and introgression in secondary contact (Fig 1B). To address these questions, we analyzed patterns of variation in our genome-wide sampling of nuclear SNPs from the hybrid and parental populations, and interpreted these using existing snake genomic resources. To explicitly test a null model of neutral drift across loci contributing to divergence between parental populations, we implement a newly developed method, *GppFst* (Adams et al. 2016), to identify patterns of allelic differentiation that are poorly explained by neutral processes and more likely driven by selection. We used analyses of genomic clines implemented in *bgc* (Gompert and Buerkle 2012) to study patterns of introgression in hybrids and identify genomic regions that resist introgression due to selection. This genomic cline approach has been used in several recent studies to explore

patterns of introgression in hybrids (e.g., (Gompert et al. 2012a; Janousek et al. 2012; Nosil et al. 2012; Parchman et al. 2013; Trier et al. 2014; Caseys et al. 2015), and simulations have demonstrated that it is robust to the detection of loci with clines that deviate from the genome-wide expectation (Gompert and Buerkle 2011). Finally, to explore potential selective pressures driving patterns of divergence and introgression, we test the hypothesis that *a priori* sets of candidate genes are enriched for outlier loci (Fig. 1B). *A priori* candidate gene sets were curated specifically to address two chief interests: i) the observed asymmetry in complements of nuclear and mitochondrial alleles in admixed individuals that may be due to incompatibility between mitochondrial haplotypes and nuclear genomic background, and ii) prominent phenotypic and natural history differences between populations of *C. atrox* (i.e., venom composition, coloration, and reproductive biology).

Materials and methods

Sampling design, RADseq data generation, and computational analysis

In a previous study (Schield et al. 2015), we generated restriction-site associated DNA sequencing (RADseq) data for 43 samples from throughout the range of *C. atrox*. To increase sampling, we generated new RADseq data from 32 additional individuals including: 15 samples west of the Continental Divide, 7 samples east of the west Texas mountains, and 10 samples from the Chihuahuan region (Fig. 1A, Supplementary Online Table 1). Sampling of these populations enabled the comparative approach illustrated in Fig. 1B. We also generated RADseq data for six individuals of the Mohave Rattlesnake (*C. scutulatus*; Supplementary Online Table 1), a close relative of *C. atrox*. These were used as the outgroup comparison for lineage delimitation analyses (see below). New data generation followed Schield et al. (2015), which

used a slightly-modified version of the double digest RADseq approach of Peterson et al. (2012), targeting ~20,000 genomic loci per individual (see Appendix I for detailed methods). We removed PCR clones from raw sequencing reads using the Stacks clone_filter module (Catchen et al. 2013), and trimmed the adapter sequence and UMI bases from filtered reads (see Appendix I). Reads were demulitplexed to individuals using the Stacks process_radtags module. Trimmed paired reads were assembled into a consensus 'pseudo-reference' genome using the *de novo* assembly option in dDocent (Puritz et al. 2014) to form contigs, using program defaults. This yielded 24,115 *de novo C. atrox* loci with an average length of 192 bases (consistent with our paired-read length). Individual read files were mapped to the pseudo-reference using the BWA-mem algorithm (Li and Durbin 2009), specifying paired read inputs, an open gap penalty of 3, and a mismatch penalty of 2. SAMtools v1.3 and BCFtools v1.3 (Li et al. 2009) were used to process mapping files and to call single nucleotide polymorphisms (SNPs). Variants were filtered using VCFtools (Danecek et al. 2011) to retain SNPs with quality scores above 30; to reduce linkage among variant sites, we sampled one variant site from each locus. We also filtered to remove SNPs with minor allele frequencies below 0.05, as these lack useful ancestry information for analyses of differentiation and introgression (Gompert et al. 2012a). We performed analyses of genes in proximity (i.e., genetically linked) to SNPs by inferring putative orthology between our pseudo-genome contigs and the King Cobra genome (Vonk et al. 2013) using a BLAST search (Altschul et al. 1990). BLAST results were filtered by e-value, and a single alignment with the lowest e-value per variable contig was retained. Here, we assumed that genetic variation consistent with natural selection observed at RADseq loci is due to physical linkage to genic targets of selection. We used the King Cobra genome as a reference because it is currently the closest relative to *C. atrox* with a fully-annotated genome sequence.

Two-dimensional allele frequency spectra and competing models of divergence

We tested competing population genetic models of divergence between parental populations using δaδi to analyze two-dimensional allele frequency spectra (2D-AFS; (Gutenkunst et al. 2009), and further investigated evidence for the presence of two distinct ancestral lineages using Bayesian lineage delimitation implemented in BFD* (Leache et al. 2014). Rather than the pipeline detailed above for downstream divergence and introgression analyses, here we used the specifications detailed in Schield et al. (2015) to generate input data because we leveraged existing workflow (Portik et al. In Review); https://github.com/dportik/dadi_pipeline) that uses Stacks (Catchen et al. 2013) output to format data for analyses in δaδi. In brief, loci were filtered in Stacks to retain loci only if they were present in both parental populations and represented by a minimum of 30% of individuals at a minimum coverage depth of 6x. This pipeline yielded 3,756 loci for analyses of demographic models and lineage delimitation.

Using δaδi, the folded 2D-AFS was generated from the nuclear SNP data for a maximum of 23 individuals from the western population and 16 individuals from the eastern population. We excluded admixed individuals and individuals near the hybrid region to infer long-term gene flow, rather than the effect of contemporary hybridization. To account for missing data we projected down to smaller sample sizes (western: 24 alleles, eastern: 24 alleles), resulting in 2,646 segregating sites. Nine alternative demographic models were fit to the 2D-AFS (Table 1, Supplementary Online Fig. 1), including models with and without divergence. For each model, 20 sets of randomly perturbed parameters were optimized by the Nelder-Mead method for a maximum of 100 iterations. The 2D-AFS was simulated from each parameter set, and extrapolation was performed with a grid size of [40,50,60]. Log-likelihoods were estimated using

the multinomial approach, and models were evaluated using the Akaike information criterion (AIC) based on the replicate with the highest log-likelihood score.

We explicitly tested for the presence of two distinct *C. atrox* lineages using genome-wide Bayes Factor lineage delimitation implemented in BFD* (Leache et al. 2014). Similar to demographic model analyses, we tested two competing hypotheses of either two divergent parental lineages (east and west lineages, Model A) or a single combined *C. atrox* lineage (Model B), using *C. scutulatus* as an outgroup (see Sampling Design). We generated a combined alignment of 8 *C. atrox* samples randomly sampled from both the eastern (n = 4) and western (n = 4) parental populations and an additional 4 outgroup samples (*C. scutulatus*) in Stacks, using the same specifications as above for δaδi input, which yielded a total alignment of 463 biallelic SNPs. We set a diffuse gamma prior ($\alpha = 1, \beta = 10$) for the population parameters (θ) based on the average pairwise genetic distance ($average\ \pi$= 0.12) across all loci, and a diffuse gamma prior ($\alpha = 1, \beta = 23.8$) on the speciation rate (λ) to reflect an expected genetic divergence between *C. atrox* and *C. scutulatus* inferred from previous studies ($E[divergence] = 0.35$, (Reyes-Velasco et al. 2013)). Additionally, we fixed the backward and forward mutation rates equal to the observed frequencies within our dataset ($\mu = 2.85, \upsilon = 0.61$). We ran path sampling analyses for 48 steps, specifying $\alpha = 0.3$, each for a total of 100,000 MCMC generations and set the burnin threshold to remove the first 10,000 generations.

We used STRUCTURE (Pritchard et al. 2000) to estimate admixture proportions across the SNP dataset. We ran STRUCTURE on nuclear SNPs under models of K = 1-10 (3 iterations each; 100,000 burnin generations, 900,000 sampled generations). We used the ΔK method (Evanno et al. 2005) implemented in StructureHarvester (Earl and Vonholdt 2012) to determine the most

likely number of K ancestral populations and visualized the Q-matrix (posterior probabilities of ancestry proportions) spatially using the tessplot function (Jay et al. 2012) in R (R Core Team 2017).

Analysis of population genetic differentiation

We used three measures to summarize nucleotide diversity and to estimate genetic differentiation among loci between eastern and western parental populations. To understand the relative allelic diversity within each population, we first estimated Nei and Li's (1979) π across loci and then calculated the difference between $\pi_{eastern}$ and $\pi_{western}$ (referred to as $\Delta\pi_{eastwest}$ hereafter). Our expectation for $\Delta\pi_{eastwest}$ is that a locus with low diversity in both populations will yield a value near 0, while a locus with greater diversity in the eastern population than western will yield a positive value, and one with greater diversity in the western population than the eastern will return a negative value. We calculated relative population genetic differentiation between populations using Weir and Cockerham's F_{ST} (Weir and Cockerham 1984). F_{ST} is a relative measure of divergence between populations, and may be influenced by differences in nucleotide diversity within populations (Cruickshank and Hahn 2014), therefore we also used an absolute measure of population differentiation, d_{xy}, to characterize divergence between eastern and western populations, using the calculation described in (Irwin et al. 2016). Weir and Cockerham's correction sometimes yields negative values of F_{ST}, and we converted any negative estimates to zero (the lower bound indicative of panmixia at a given locus).

We designated outlier loci with greater than expected values of population differentiation using a multivariate approach, in the program MINOTAUR (Verity et al. 2016), using information from distributions of F_{ST}, d_{xy}, and $\Delta\pi_{eastwest}$ to identify loci that differ significantly from the genomic

background. An advantage of this approach is that it makes no specific assumptions based on a single summary statistic or the demographic history of the populations, but rather uses the combined information provided by multiple statistics and is robust for detecting genomic outliers derived from a wide variety of evolutionary scenarios. We specifically used the Mahalanobis Distance (MD, defined as the distance from the multivariate centroid; (Mahalanobis 1936) measure to delineate outlier loci where differentiation between parental populations has likely been driven by selection. Specifically, loci were considered statistical outliers if the locus-specific MD exceeded the 95th quantile of the genomic distribution, and were considered strong outliers if they exceeded the 99th quantile.

We also wished to determine if statistical outlier loci had estimates of allelic differentiation that would not be expected under a model of neutral evolution. To test this, we designed a software program, *GppFst* (Adams et al. 2016); https://github.com/radamsRHA/GppFst), which conducts posterior predictive simulations (PPS) of F_{ST} and d_{xy} under a strict model of evolution in the absence of selection. To conduct PPS analysis, we estimated divergence time and population parameters for the parental eastern and western populations using the 7,031 SNPs detailed above ($\tau_{eastwest}$, θ_{west}, θ_{east}, and $\theta_{eastwest}$) via Markov Chain Monte Carlo (MCMC) sampling implemented in the program SNAPP (Bryant et al. 2012). We stress that parameterization in SNAPP for *GppFst* was done using only structured parental populations and did not include admixed individuals. Additionally, we found very strong Bayes Factor (BF) support for two distinct *C. atrox* lineages (BF = -336; see Results); we would not expect strong support for this model in the BFD* framework if gene flow were pervasive between the eastern and western populations (Zhang et al. 2011). The mean posterior estimates of population genetic parameters used for *GppFst* simulations from SNAPP were as follows: $\theta_{east} = 0.217$, $\theta_{west} = 0.0592$, $\theta_{eastwest} = 0.0615$,

and $\tau_{eastwest}$ = 0.00578. We ran MCMC for a total of 500,000 generations, sampling every 1,000 generations. We assessed posterior convergence and stationarity using Tracer (for all parameters ESS > 500; (Drummond and Rambaut 2007) and discarded the first 125,000 (25%) steps as burn-in, leaving total of 375 MCMC samples used to generate a PPS F_{ST} distribution. For each MCMC step, we then simulated 100 independent loci with a length of 190 base pairs (equal to the average length of variant loci) under a JC69 model using the R package phybase (Liu and Yu 2010) with random sampling of individuals from the empirical distributions of locus coverage for eastern and western samples, respectively.

We calculated F_{ST} and d_{xy} for a randomly sampled polymorphic site for each simulated locus to generate a theoretical distribution of values, which allowed us to compare empirical distributions of F_{ST} and d_{xy} to identify loci with levels of divergence that are poorly explained by the neutral model of divergence. Specifically, we calculated the probability that the proportion of empirical loci with a given value of allelic differentiation is observed in the posterior predicted distribution, and thus were able to reject a strict model of neutral evolution where this probability is very low. Additionally, the use of simulations allowed us to account for the potential influence of unevenness in sample sizes across loci and differences in effective population size between the two populations. We considered statistical outliers from multivariate outlier detection 'positives' for selection if they were also poorly explained by a neutral model of evolution based on their comparison to simulated F_{ST} and d_{xy} distributions in *GppFst*.

Analysis of genomic introgression

We conducted Bayesian estimation of genomic clines using the program *bgc* (Gompert and Buerkle 2011) to identify loci that defy neutral expectations of admixture compared to the

genomic background. We segregated samples into western, eastern, and admixed populations (Fig. 1), and calculated locus-specific allele frequencies for each population. *bgc* estimates the hybrid index (h) for each individual in an admixed population, representing the proportion of an individual's genome inherited from one parental population. Hybrid index and two locus-specific genomic cline parameters (α, the genomic cline center parameter; and β, the genomic cline rate parameter) are used to estimate the posterior probability of inheritance from one parental population (Φ) at a given locus within the admixed population. Under this model, if both α and β are equal to zero, h and Φ will be equivalent (Gompert and Buerkle 2011), and will match the neutral genomic background expectation (dashed line in the genomic cline panel of the 'study design' in Fig. 1B). Loci enriched for selection can be identified based on each cline parameter and classified based on their locus-specific cline parameters relative to a genome-wide distribution (Gompert and Buerkle 2011; Gompert et al. 2012a).

We used two methods to identify outlier loci that exhibit exceptional introgression compared to the neutral genomic background expectation. First, we considered loci outliers if they had evidence of excess ancestry from one parental population (i.e., the 95% confidence interval of the α parameter did not include 0; see (Gompert and Buerkle 2012) and also had a median α within the tails of the 95th quantile of the median α distribution. Second, 'strong' statistical outliers were determined using the locus-effect quantiles for α (γ-quantile) relative to a genome-wide distribution (Gompert and Buerkle 2011); loci were considered strong outliers if their γ -quantile fell outside of the interval q_n, where $\frac{1-n}{2} < q_n < \frac{n}{2}$, where $n = 0.1$. While both statistical outlier loci (from divergence analyses) and loci with evidence of excess ancestry are not definitively targets of selection, they do exhibit patterns that are poorly explained by neutral

introgression and are often found in genomic regions impacted by strong selection (Gompert and Buerkle 2012).

To compare empirical genomic clines to expectations of neutral introgression, we simulated a 'null' admixed population by randomly sampling alleles observed in the empirical parental populations for each locus equal to the number of individuals in the empirical admixed population ($n = 21$). Here allelic information per individual per locus was derived from two random draws from parental alleles (eastern or western) for that locus. To evaluate multiple random draws of parental alleles for each locus per individual, we repeated this process five times and ran parallel *bgc* analyses to establish a benchmark for expectations under neutral introgression using the specifications detailed in Appendix I.

Comparative analyses of divergence and introgression

We tested the hypothesis that there is a genome-wide relationship between allelic differentiation accumulated during divergence and patterns of introgression in secondary contact using linear correlation analyses. We compared locus-specific univariate nucleotide diversity and allelic differentiation statistics ($\Delta\pi_{eastwest}$, F_{ST}, and d_{xy}) as well as the multivariate Mahalanobis Distance from divergence analysis with the genomic cline center parameter, α. We examined these relationships using the absolute values of $\Delta\pi_{eastwest}$ and α.

We tested for significant overlap in genes linked to outliers from divergence and introgression analyses and for enrichment of candidate gene sets related to traits of interest by first determining the nearest up- and down-stream annotated genes from each orthologous cobra genome region. We then built a distribution of quantities of overlapping genes based on random expectation by producing 1,000 randomly resampled datasets using the background list of all

genes linked to sampled loci, and compared these to observed quantities of overlapping genes between outliers from divergence and introgression analyses.

We specified five *a priori* candidate gene sets of interest: snake venom genes, coloration genes, genes involved in reproductive output and timing, nuclear-encoded mitochondrial proteins (nuc-mt), and nuclear-encoded subunits specifically involved in oxidative phosphorylation (nuc-oxphos), and tested for enrichment of these candidate gene sets within the sets of genes we identified as being linked to outlier loci from divergence or introgression analyses using Fisher's Exact Tests. Venom, reproduction, and coloration candidate gene sets were chosen to enable us to explicitly test if genes related to known phenotypic differences between *C. atrox* populations are highly differentiated and also contribute to reproductive isolation in hybrids. Nuc-mt and nuc-oxphos gene sets were curated to test for patterns of enrichment in divergence and introgression outliers because of their interactions and co-evolution with the mitochondrial genome. Here, we were specifically interested if a decoupling of this co-evolution due to introgression and incompatibility between nuclear and mitochondrial genomes could explain the previously observed asymmetry in complements of nuclear alleles with western and eastern mitochondria (Schield et al. 2015). For additional rationale behind these specific candidate gene sets, see the Introduction, and see Appendix I for details of candidate gene BLAST set construction and analysis.

Results

Results of variant calling pipeline for divergence and introgression analyses

The assembled *C. atrox de novo* 'pseudo-reference genome' included 24,115 loci constructed from sequenced RADseq reads. Following mapping and filtering to remove indels, spurious base

calls introduced by sequencing error, non-biallelic sites, and multiple variants per locus, we retained 7,031 SNPs for divergence and introgression analyses (see below). Orthologous regions from blast-matching of *C. atrox* contigs to the cobra genome were fairly evenly spread across 3,832 cobra scaffolds, with a range of 1 – 20 (mean of 1.87) *de novo* contigs per cobra scaffold (cobra scaffold N50 = 226 Kb). While the proportion of missing data was variable across loci (0 – 96%), we did not observe significant correlations between missing data and allelic differentiation statistics ($p = 0.538$) or genomic cline parameter estimates ($p = 0.917$).

Evidence for divergence in isolation and secondary contact of lineages

Analysis of the 2D-AFS and competing population genetic models in δaδi indicated that models that did not include population splitting fit poorly to our data when compared to models that included lineage divergence. Untransformed parameters are provided in Table 1, along with an estimate of θ ($\theta = 4N_{ref}\mu L$, where L is the total length of sequenced region SNPs were ascertained from). A model of population divergence in isolation followed by recent gene flow in secondary contact provided the best fit our data (Fig. 2A-C, Table 1), with minor residuals. This model was very strongly supported with an Akaike weight = 0.78, which fit the data substantially better than even the second best supported model of divergence with secondary contact and asymmetric gene flow (Akaike weight = 0.15). Akaike weights were negligible for all other models indicating essentially no support for any of these alternative scenarios (see (Burnham and Anderson 2003). Under the model of divergence with secondary contact, δaδi estimated greater effective population size and genetic diversity in the eastern population than in the western, consistent with previous studies comparing demographic parameters (Castoe et al. 2007; Schield et al. 2015). The best-fit model of divergence with secondary contact also suggests equivalent

migration rates between eastern and western populations, contrary to previously inferred patterns from a single mitochondrial locus.

Results for the competing hypotheses of two distinct lineages versus a single-lineage model of *C. atrox* are shown in Table 2. We found very strong (decisive; (Kass and Raftery 1995) support for the '3 species' model, which included *C. scutulatus* as the outgroup + two parental lineages of *C. atrox*, consistent with previous mtDNA inferences (Castoe et al. 2007; Schield et al. 2015) and demographic model selection in δaδi. Collectively, these inferences support two distinct ancestral/parental lineages of *C. atrox*. STRUCTURE analysis yielded a best-fit model of $K = 2$ ancestral populations, which was supported using the ΔK method ($\Delta K[2] = 192.7$, while $\Delta K[3] = 129.3$; (Evanno et al. 2005). We observed high posterior probability assignments to each ancestral population cluster in regions outside of the Chihuahuan region, and a gradient of mixed assignments within this region (Supplementary Fig. 2). We inferred the greatest degree of admixture (i.e., Q-values near 0.5) within samples near the Continental Divide and in western Texas.

Patterns of genetic differentiation between parental lineages

Locus-specific estimates of genetic diversity and differentiation between parental eastern and western populations varied considerably (Fig 3). Most loci exhibited low relative differentiation between populations (mean $F_{ST} = 0.09$, median $F_{ST} = 0.022$), however the range of estimates included a number of SNPs with high F_{ST} values (Fig 3A), including 10 SNPs (0.14%) with fixed differences (i.e., $F_{ST} = 1$). We also observed 28 SNPs (0.39%) with F_{ST} estimates greater than 0.8, and 216 (3.07%) greater than 0.5 – values considered consistent with a high degree of population differentiation (Hartl and Clark 1997). The distribution of absolute allelic

differentiation, d_{xy}, between parental populations was generally lower, as expected, but largely consistent with the F_{ST} distribution (Fig 3B). While there was not a perfect linear relationship between genome-wide estimates of these two parameters per locus ($r = 0.58$), loci with extreme F_{ST} tended to also have extreme values of d_{xy}. For example, loci with fixed differences estimated using F_{ST} also had fixed absolute differentiation ($d_{xy} = 1$). Relative nucleotide diversity ($\Delta\pi_{eastwest}$) ranged from -1 to 1, however estimates near zero were observed much more frequently (mean = 0.055, median = 0.041), and the 95% confidence interval of $\Delta\pi_{eastwest}$ ranged from -0.42 to 0.48.

We detected 299 statistical outliers and 60 strong statistical outliers that exceeded the 95th and 99th quantiles of the multivariate Mahalanobis Distance (MD), respectively. Statistical outliers were associated with elevated estimates of allelic differentiation relative to the genome-wide distribution (i.e., minimum $F_{ST} = 0.16$, minimum $d_{xy} = 0.11$), and strong statistical outliers had a minimum $F_{ST} = 0.27$ and a minimum $d_{xy} = 0.11$. We note that some loci with no evidence of allelic differentiation between populations had high MD values; loci could exhibit this pattern due to balancing selection, where shared intermediate frequency alleles are maintained by selection and are expected to result in low F_{ST} estimates. Given our inability to determine this explicitly, we removed such loci (with low F_{ST} estimates) from comparative analyses of outlier loci.

To examine the fit of our empirical data to purely neutral expectations, we binned empirical F_{ST} and d_{xy} values above discrete ranges of 0.1 (i.e., 0.0 – 0.1, 0.1 – 0.2, etc.) and compared relative frequencies of empirically observed values and those from simulations conducted using *GppFst* (Supplementary Online Fig. 3). Because results were qualitatively similar for both absolute and

relative allelic differentiation, we report here on F_{ST} results only. We found that the empirical data included far more extreme F_{ST} values (e.g., $F_{ST} > 0.5$) than that expected based on the simulated dataset (Fisher's Exact $p < 1 \times 10^{-15}$); significantly higher proportions of empirical values were observed at all intervals above $F_{ST} = 0.2$, and this was especially pronounced at very extreme F_{ST} values. For example, the empirical dataset had 10-fold more loci than expected with an $F_{ST} = 1$, which was highly significant compared to expectations derived from the neutral model (computed from simulated distribution in *GppFst*; $p < 0.0001$). Thus, based on comparisons of empirical allelic differentiation and neutral expectations, we consider that proportions of outlier loci with $F_{ST} > 0.2$ are poorly explained by neutral evolution and have more likely been driven by selection. We therefore retained 278 statistical outlier loci that exceeded the $F_{ST} = 0.2$ threshold. All 60 strong statistical outliers met this criterion, and were used in downstream candidate gene tests. Importantly, *GppFst* simulations suggest that, at a locus-specific scale, even strong outliers likely contain false positives. Specifically, based on our *GppFst* simulations, we estimate that there are roughly equal proportions of true and false positives when all loci with $0.2 < F_{ST} < 1$ are considered. The proportion of false positives substantially decreased, however, with more extreme values of F_{ST} (e.g., 19% estimated false positives at $0.6 < F_{ST} < 1$).

Patterns of genomic introgression

Locus-specific introgression also varied widely within the admixed *C. atrox* population (Fig. 3A). While most loci exhibited clines consistent with the expectation of neutral introgression between parental populations (i.e., nearly equal hybrid index and probability of parental ancestry), some locus-specific clines deviated from this pattern considerably. Estimates of the genomic cline center parameter, α, ranged from -1.66 to 1.7. The average hybrid index (genomic

background of eastern versus western ancestry) of individuals sampled from the admixed population was 0.462, consistent with a nearly completely admixed genomic background (Fig. 3A). Given the hybrid index of an individual, particular loci may deviate from the background because they are dominated by either eastern or western alleles, as indicated by the value of α. Locus estimates of the cline rate parameter, β, were less variable, ranging from -0.011 – 0.012, and the 95% confidence interval for all estimates included zero. Locus-specific clines were thus estimated predominately based on α, and calculated using a β parameter that did not deviate significantly from zero. We identified 283 outlier loci with evidence of excess ancestry (4.02% of loci) based on the distribution of the α parameter. Of these, 133 loci had evidence of western ancestry and 150 loci had evidence of excess eastern ancestry. Despite a greater number of loci with excess eastern ancestry, this difference was not statistically significant ($p = 0.337$), indicating that there was no evidence of any strong genome-wide directional preference towards alleles from one parental lineage over the other in hybrid individuals. We identified 113 strong statistical outliers using locus-specific γ-quantiles, where n equals 0.1 ($\gamma_{0.1}$ outliers; see Methods). We found more statistical outliers with eastern ancestry (70 loci) than western (43 loci), and this difference between the number of eastern and western strong statistical outliers was significant ($p = 0.0137$).

The results of genomic cline analyses on simulated admixed population data were consistent with expectations of neutral introgression (Supplementary Online Fig. 4; see also (Gompert and Buerkle 2011). Under these simulations, the probability of parental ancestry was not expected to deviate significantly from predictions based on the hybrid index. In line with this expectation, we observed zero loci with evidence of excess ancestry and no loci were identified as statistical

outliers based on the criteria used for empirical analyses; the most extreme values of α were an order of magnitude smaller than in empirical analyses (max. $\alpha = 0.18$, min. $\alpha = -0.13$).

Comparative analyses of genomic introgression and divergence

We found positive relationships between locus-specific measures of allelic differentiation and nucleotide diversity and the genomic cline center parameter α (Fig. 4B-D), with linear correlations that were statistically significant. Specifically, the correlations of $|\alpha|$ with F_{ST}, d_{xy}, and $|\Delta\pi_{eastwest}|$ were: $r = 0.22$ ($p = 2.2 \times 10^{-16}$), $r = 0.12$ ($p = 2.2 \times 10^{-16}$), and $r = 0.22$ ($p = 2.2 \times 10^{-16}$), respectively. We also found a positive relationship between the multivariate Mahalanobis Distance from divergence statistics and $|\alpha|$ ($r = 0.08$, $p = 2.61 \times 10^{-9}$). Because we found significant, positive correlations between all divergence estimates and introgression, we report further here on F_{ST} only. In addition to the overall genomic trend, we observed positive correlations when subsets of loci were considered (e.g., loci with $F_{ST} > 0.2$, $F_{ST} > 0.5$, etc.; Supplementary Fig. 5). There were a total of 28 loci that were outliers based on both allelic differentiation and excess ancestry in introgression. These contained similar proportions of positive (eastern ancestry; 53.6%) and negative (western ancestry; 46.4%) α values; these proportions were not significantly different ($p = 0.82$). The same set of loci contained 10 $\gamma_{0.1}$ outliers with eastern origin and 3 with western origin. Interestingly, however, the most extreme (e.g., top 10) statistical outliers from divergence and introgression analyses were non-overlapping (unique outlier loci).

Comparative analyses of gene overlap and candidate gene sets

There was higher than expected overlap in genes putatively linked to loci from divergence and introgression analyses when we compared outliers and strong outliers from these analyses. In

both cases, the gene overlap in the empirical dataset exceeded the 95th quantile of overlap observed in resampled datasets (Supplementary Online Fig. 6A-B). It is notable, however, that this is the reverse for the 10 most extreme outliers from each analysis, which did not overlap at all.

Analyses of *a priori* candidate genes revealed differential patterns of enrichment between sets of genes linked to divergence and introgression outliers (Fig. 5, Supplementary Online Table 2). Venom genes were enriched in gene sets from both divergence outliers ($p = 0.0051$) and excess ancestry outliers from introgression ($p = 7.62 \times 10^{-5}$), but not more stringent divergence and $\gamma_{0.1}$ outliers, possibly due to low statistical power from small sample sizes (Fig. 5A-C). Reproduction candidate genes were enriched in outliers from allelic differentiation ($p = 0.0163$), but not introgression (Fig. 5D-F). There was no evidence for coloration gene enrichment; a single coloration gene was found among excess ancestry outliers, and none were included with divergence outliers. Nuc-mt and nuc-oxphos candidate genes exhibited the opposite pattern of reproduction genes (Fig 5G-L), with evidence of enrichment in genes from introgression analysis but no evidence of enrichment in outliers from divergence. Nuc-mt genes were enriched in both excess ancestry ($p = 0.018$) and $\gamma_{0.1}$ introgression outlier genes ($p = 0.0097$), and nuc-oxphos genes were enriched in excess ancestry outliers ($p = 0.014$). Nuc-oxphos genes with excess ancestry included several ATPases and a nuclear subunit directly involved in oxidative phosphorylation. Specifically, the genomic cline center parameter estimate for the locus putatively linked to *UQCRFS1* (encoding cytochrome c reductase), a component of the cytochrome b-c1 complex (complex III), was consistent with western ancestry ($\alpha = -1.071$, 95% CI = -1.792 – -0.302; Fig. 5K).

The relationship between measures of allelic differentiation and introgression for candidate gene sets also varied, and in some cases differed considerably from the overall genomic pattern. F_{ST} and $|\alpha|$ were negatively correlated for venom genes (r = -0.53, p = 0.00023; Fig. 5C). In contrast, there was a positive correlation between F_{ST} and $|\alpha|$ for reproduction genes (r = 0.62, p = 0.032; Fig. 5F), which was somewhat surprising given a lack of evidence for enrichment for reproduction in introgression outliers. Finally, we did not find significant correlations between F_{ST} and $|\alpha|$ for the nuc-mt and nuc-oxphos candidate gene sets, but the trend was towards lower F_{ST} distributions (Fig. 5G, J) associated with high values of α, which is consistent with evidence for enrichment only in introgression.

Discussion

Maintenance of lineage integrity upon secondary contact

In this study, we found evidence consistent with previous inferences that populations of *C. atrox* were historically isolated and have more recently experienced secondary contact with hybridization (Tables 1 and 2; Fig. 2; Supplementary Online Fig. 2). The homogenizing effects of pervasive gene flow upon secondary contact should in theory result in a reversal of the divergence and speciation process (Taylor et al. 2006), however it is becoming clear from population genomic data from natural systems that this is not always the case. Instead, while much of the genome introgresses freely in hybrids, certain regions may resist gene flow due to selection for or against incoming parental alleles because of a direct or indirect impact on hybrid fitness. In the case of *C. atrox*, the hybrid zone might be considered a 'tension zone,' where a potentially small proportion of the genome acts to limit homogenization of lineages through partial reproductive isolation (Gay et al. 2008). A comparative dissection of the patterns of allelic

variation involved in lineage divergence and lineage convergence in secondary contact may thus provide key insight for reconciling interactions between these processes in speciation, and particularly how selection and stochastic forces shape the relationship between these processes. Further identification of common loci under selection in both scenarios implies that there may be links between local adaptation and the maintenance of lineage integrity upon secondary contact, and that traits locally co-adapted in allopatric divergence may also contribute to reproductive isolation. It is poorly understood to what degree we expect loci to be involved in both processes, however, as diverse aspects of demography, time since divergence, geographically diverse selection pressures, and natural history are likely to be influential. Here, we have explored this relationship in Western Diamondback Rattlesnakes, and discuss below ways in which adaptive evolution in divergence and introgression appear to be both linked and also idiosyncratic.

Evidence for selection in divergence and introgression

To address the major hypotheses in this paper, we were interested in identifying loci with patterns likely driven by selection to determine if adaptation in divergence and fitness in hybrids are linked. We find a highly heterogeneous genomic landscape of allelic differentiation between parental *C. atrox* populations, similar to that reported in recent studies in other species (Ellegren et al. 2012; Nadeau et al. 2012; Flaxman et al. 2013; Ferchaud and Hansen 2016). This distribution included loci with extreme allelic differentiation that are likely linked to genomic regions under selection. Because extreme differentiation could evolve due to stochastic (i.e., drift) or deterministic processes, we tested for signatures of selection-driven evolution, and employed a novel simulation method (*GppFst*; (Adams et al. 2016) to compare empirical distributions of allelic differentiation between parental populations to distributions expected under a neutral coalescent model (including mutation and drift). We found extreme divergence in

neutral simulations to be comparatively rare relative to the empirical distribution, suggesting that outlier loci are enriched for loci under selection (Supplementary Online Fig. 3).

Even setting cutoffs to highly stringent levels, our simulations suggest that statistical outlier loci are expected to contain false positives, and locus-specific conclusions should be interpreted with appropriate caution. We note that evolutionary forces not explicitly explored here (e.g., recombination) could also drive false inferences of selection. For example, recombination hotspots may mimic patterns of extreme differentiation due to selection (Myers et al. 2005), especially if high recombination rates are paired with rapid mutation rates or biased gene conversion, as has been observed in humans (Lachance and Tishkoff 2014) and *Drosophila* (Kulathinal et al. 2008). Alternatively, low recombination paired with background selection or neutral divergence could drive differentiation between populations (O'Reilly et al. 2008). With a single pairwise comparison of two populations/lineages in this study, our ability to explicitly account for recombination is limited. However, evidence for biological links between inferred selected loci and enrichment of *a priori* candidate genes underlying major differences between *C. atrox* populations argues that false positives have not substantially hampered our ability to detect patterns of variation that we expect have been driven by adaptation. Nonetheless, further investigation using information from multiple lineage pairs would be valuable for distinguishing between patterns driven by recombination and 'selective sweeps'.

As in divergence analyses, we found the genome-wide landscape of introgression in hybrid *C. atrox* to vary considerably, containing extreme locus-specific clines far outside the expectations of neutral introgression (Supplementary Online Fig. 4). For these loci, we propose that selection has acted to resist gene flow to either maintain or resist alleles from one parental population

more than would be expected based on the presumably neutral background genomic pattern (Fig. 4A; (Gompert et al. 2012b). Loci may fit these extreme patterns because alleles from one parental population offer an increase in hybrid fitness (Payseur 2010), or because alleles from one parental population are selected against because they are incompatible with the opposite parental genomic background (i.e., Dobzhansky-Muller incompatibilities; (Dobzhansky 1936; Muller 1942; Orr 1995).

Genomic relationship between divergence and introgression in secondary contact

The comparison of allelic variation in divergence and introgression revealed an intriguingly complex association across the genome, and broadly identified genomic regions with elevated genetic differentiation between parental *C. atrox* populations that also appear to be under selection in hybrids. Overall, this relationship supports the hypothesis that selection-driven genomic changes in divergence may also play a role in reproductive isolation in secondary contact, and may further be central to understanding why incipient species may not completely homogenize when they enter secondary contact. This comparison allows us to quantify how tightly linked divergence and admixture processes may be in natural systems, and how divergence in turn translates to reproductive isolation.

A positive relationship between locus-specific measures of allelic differentiation and the absolute value of α has also been reported in recent studies (Gompert et al. 2012a; Nosil et al. 2012; Parchman et al. 2013). Our findings certainly agree that this relationship exists and provide novel insights into this association in several ways. First, previous studies have specifically considered patterns of variation in *Lycaeides* butterflies (Gompert et al. 2012a), *Timema* stick insects (Nosil et al. 2012), and *Manacus* birds (Parchman et al. 2013). The rattlesnake system studied here is

distinct in several respects, including its expansive geographic range encompassing diverse habitat and climatic conditions, dispersal capabilities (i.e., flying birds and insects [*Lycaeides*, specifically] versus crawling reptiles), life history traits, and inferred levels of divergence between parental populations. Given biological differences relevant to divergence and introgression between rattlesnakes and previously studied taxa, it is remarkable that this relationship is observed consistently across these disparate systems.

Second, It is notable that the association between allelic differentiation and introgression in *C. atrox* was less than would be expected if F_{ST} and α were tightly biologically autocorrelated (Fig 4B-D). This was made apparent by the small proportion (0.4%) of loci that were strong outliers in both analyses and in the 10 most extreme outliers from each analysis not overlapping at all, despite greater than expected overlap in genes linked to outliers overall (Supplementary Online Fig. 6; see also (Nosil et al. 2012). In other words, the most extreme outliers in divergence were poor predictors of the most extreme outliers in introgression, suggesting that being under strong selection in divergence is not a singular factor driving importance in introgression. Such idiosyncratic patterns of selection were also observed in variable enrichment of candidate gene sets (see below) that revealed intriguing distinctions between *C. atrox* and previously studied systems. In particular, loci that were overlapping extreme outliers from both analyses showed remarkably symmetrical patterns of excess ancestry from eastern and western populations. This is in contrast to overlapping outliers in *Lycaeides* and *Timema*, where there was instead a substantial bias towards excess ancestry from a single parental population over the other (Gompert et al. 2012a; Nosil et al. 2012). The symmetry of excess ancestry in outliers from *C. atrox* supports the hypothesis that adaptive changes in both ancestral populations during divergence contribute to reproductive isolation in hybrids.

Biological interpretations of adaptive evolution in divergence and introgression

Our analyses provide exciting evidence linking patterns of genome-wide selection in divergence and introgression with phenotypic traits and expected patterns of prior interest in this species. Specifically, we found multiple candidate gene sets related to known phenotypic and life history differences between *C. atrox* populations, along with putative incompatibilities in hybrids, were enriched for selection. Additionally, correlations between allelic differentiation and introgression for candidate genes varied considerably and were sometimes contrary to the overall genomic relationship, further highlighting multifarious patterns of divergent selection and selection against maladaptive alleles in hybrids. For example, the strong negative correlation between divergence and introgression for venom genes (Fig. 5C), despite evidence for enrichment for venom in outliers from both analyses, suggests that specific targets of selection can be fairly unique at the locus-level between these processes. An intriguing possibility is that this pattern is driven by the extreme toxicity of venom components and the diversity of tissues and physiological functions they disrupt (Mackessy 2008), which may drive complex interactions between specific venom alleles and the genomic background to prevent or minimize auto-toxicity or other deleterious effects. This diversity likely leads to broad variation in the degree of coevolution of venom alleles with other biological systems that protect rattlesnakes from the action of their own venoms (Noguchi 1909; Nichol et al. 1933; Nahas et al. 1983). Variation in levels of venom-genome coevolution, together with geographic variation in prey-specific venom allele effectiveness (Perez et al. 1979), may explain our findings that selection on venom-linked loci is idiosyncratic between divergence and admixture despite broad enrichment of venom loci as targets of selection in both processes.

Because differences in reproductive output (i.e., size and number of oocytes and follicles) between eastern and western *C. atrox* females (Spencer 2003) could result from adaptation to local ecological constraints (Qualls 1997; Caro et al. 2009), we tested the hypothesis that candidate genes involved in reproduction were enriched for selection. Reproduction-related genes were enriched in divergence outliers, and the distribution of allelic differentiation among these loci was high overall relative to the genomic background (Fig. 5D). Though reproduction-related genes were not statistically enriched in outliers from introgression analysis, they had F_{ST} and α estimates that were tightly correlated (Fig. 5F), showing the opposite pattern observed for venom and a pattern that is predicted by the genomic relationship between divergence and introgression. Thus, analyses of reproduction gene enrichment provide indirect evidence for divergent reproductive adaptations between parental populations that may also have an effect on hybrid fitness.

Of the candidate gene sets we considered, genes underlying vertebrate coloration and patterning were the only set with no evidence of enrichment in either divergence or introgression. Indeed, we only observed a single 'coloration' gene among outlier loci, which had evidence of excess eastern ancestry. We considered multiple possible reasons for this result. The first is that the detection of signatures of adaptation across the genome using RADseq is limited by a small representative sample of loci (Lowry et al. 2016), and thus we may have lacked power to detect these selection in these genes. Alternative explanations include that neutral processes have shaped differences in coloration between populations of *C. atrox*, or that selection on coloration is highly locally adapted on a very fine geographic scale that our sampling and analyses were not designed to detect (Klauber 1956; Sweet 1985; Campbell and Lamar 2004; Farallo and Forstner 2012).

Evidence for selection against introgression of nuclear-encoded genes that co-evolve with the mitochondria (i.e., nuc-mt and nuc-oxphos genes) supports inferences from our previous studies that found asymmetry in complements of nuclear allelic content paired with eastern and western mitochondria, suggesting this observation is likely due to cytonuclear incompatibilities (Schield et al. 2015). Further, the lack of evidence for divergent selection on nuc-mt and nuc-oxphos genes combined with strong selection against introgression in hybrids matches predictions of hybrid incompatibility accumulation (Orr 1995). We found exceptional introgression at a gene (*UQCRFS1*; encoding cytochrome c reductase) that interacts directly with mitochondrially-encoded subunits of the oxidative phosphorylation cascade. Theory and examples from empirical studies predict that a decoupling of the co-evolution between this gene and the mitochondrial background via introgression should impact hybrid fitness by producing maladaptive combinations of nuclear and mitochondrial alleles (Burton et al. 2006; Ellison et al. 2006; Ellison et al. 2008; Sloan et al. 2016). We found strong evidence of excess ancestry at this and other nuclear-encoded mitochondrial genes, consistent with selection to reduce the decoupling effect of introgression that appears to in turn contribute to the partial reproductive isolation of *C. atrox* lineages in secondary contact.

Conclusions

Population genomic studies of historically isolated populations that have experienced secondary contact but maintain some level of lineage integrity through partial reproductive isolation provide new insight into this seemingly paradoxical process. They also provide an important and unique indication of the potential roles of selection in the processes of speciation in allopatry, and in the context of continued maintenance of isolation and lineage integrity in secondary contact. We found compelling evidence that, while the landscape of divergence and introgression

is complex, a genome-wide relationship between these processes supports the hypothesis that divergent selection in allopatry also contributes to the maintenance of lineage integrity upon secondary contact. In addition to evidence for selection in sets of genes underlying divergent phenotypes between *C. atrox* populations, we find evidence for an important role of loci that have diverged neutrally yet lead to incompatibilities in hybrids (e.g., genes involved in oxidative phosphorylation). Further studies examining secondary contact zones between lineages, particularly lineages with varying levels of divergence, will provide valuable extensions to the work presented here to test the evolutionary replicability and generality of such patterns across species and the speciation continuum.

Acknowledgments

We thank Jesse Meik, Corey Roelke, Jeff Streicher, Hans-Werner Herrmann, Jim McGuire (MVZ) and Jens Vindum (CAS) for providing specimens and tissue loans. Support was provided by startup funds from the University of Texas Arlington to TAC, NSF Doctoral Dissertation Improvement Grants to DRS & TAC and DCC & TAC (NSF DEB-1501886, DEB-1501747), and from the Venom Viper Resource Grant (NIH/ORIP-P40ODO10960) to EES.

FIGURES

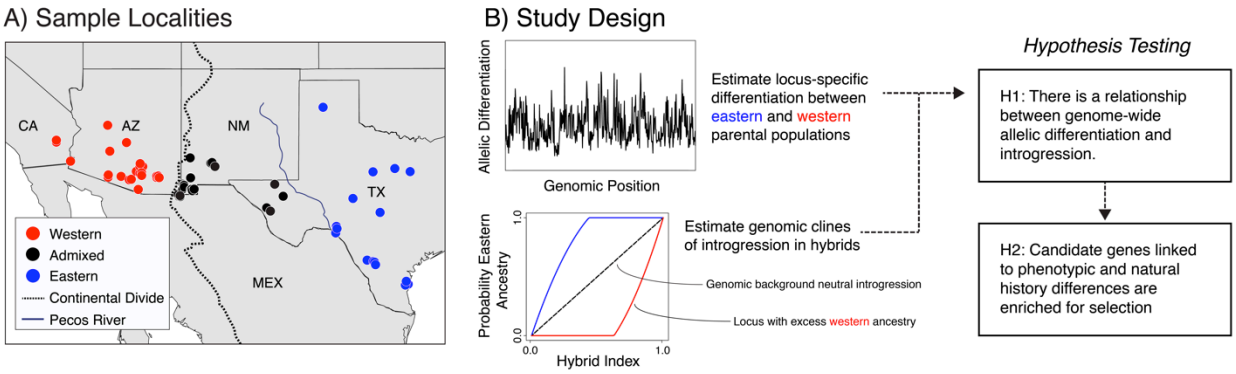


Figure 1. Sampling and study design. **A.** Map of sampled localities, where red, blue, and black circles represent western, eastern, and admixed populations, respectively. The Continental Divide is depicted as a dashed line and the blue line represents the Pecos River. **B.** Schematic view of the study design in this paper. The first steps were to estimate allelic differentiation between parental populations and genomic introgression within hybrids to identify outlier selected loci. In the lower panel, the probability of parental ancestry is compared to the hybrid index of the admixed population. Here, the dashed line represents the expectation of neutral introgression. The blue and red lines represent loci with evidence of excess eastern and western ancestry, respectively, due to the action of selection. Allelic differentiation and introgression estimates were then compared to determine if a genome-wide relationship between divergence in allopatry and admixture upon secondary contact exists, and we tested if genes linked to outlier loci were enriched for candidate gene sets of interest.

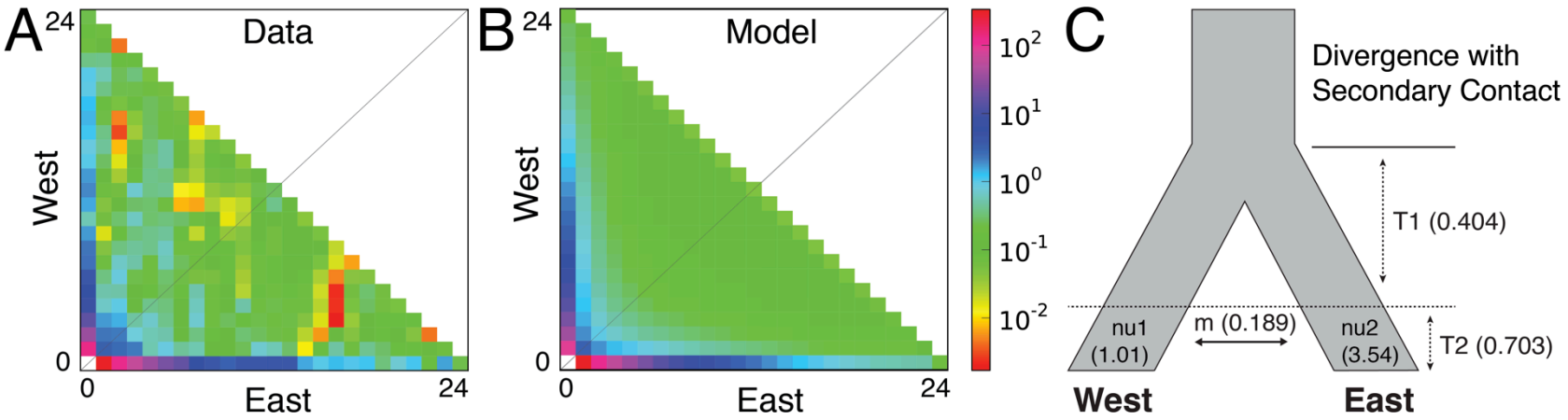


Figure 2. Summary of the two-dimensional allele frequency spectrum (2D-AFS) between parental populations and results of demographic model testing. **A-B**. 2D-AFS plots of empirical data and the inferred best-fit model predictions, respectively. A legend of the spectrum of allele frequencies is show to the right of B. **C.** Inferred best-fit model of population divergence in isolation followed by recent secondary contact, with demographic parameter estimates from δaδi. For details of parameter estimates and abbreviations, see Table 1.

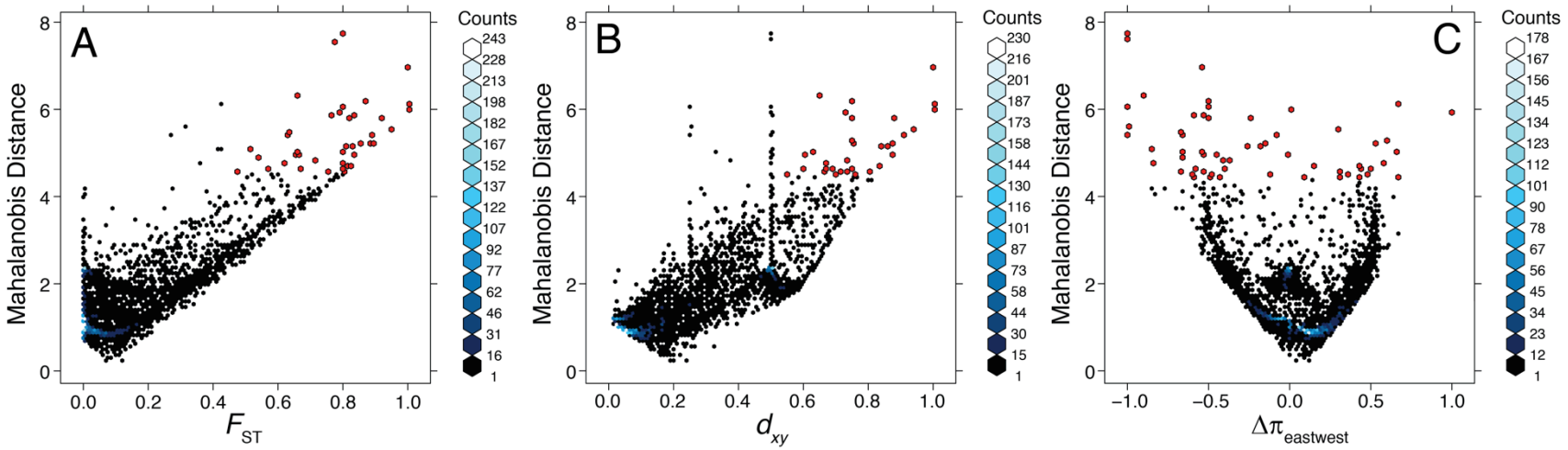


Figure 3. Variation in allelic differentiation between parental eastern and western populations of *C. atrox*. Each comparison depicts the distribution of one univariate measure of differentiation compared to the locus-specific Mahalanobis distance. **A.** Relative differentiation, F_{ST}. **B.** Absolute differentiation, d_{xy}. **C.** Relative nucleotide diversity between eastern and western populations, $\Delta\pi_{eastwest}$. Scales to the right of each panel represent the distribution of bins that quantities of loci fell into within each distribution. Large, red points depict strong statistical outliers.

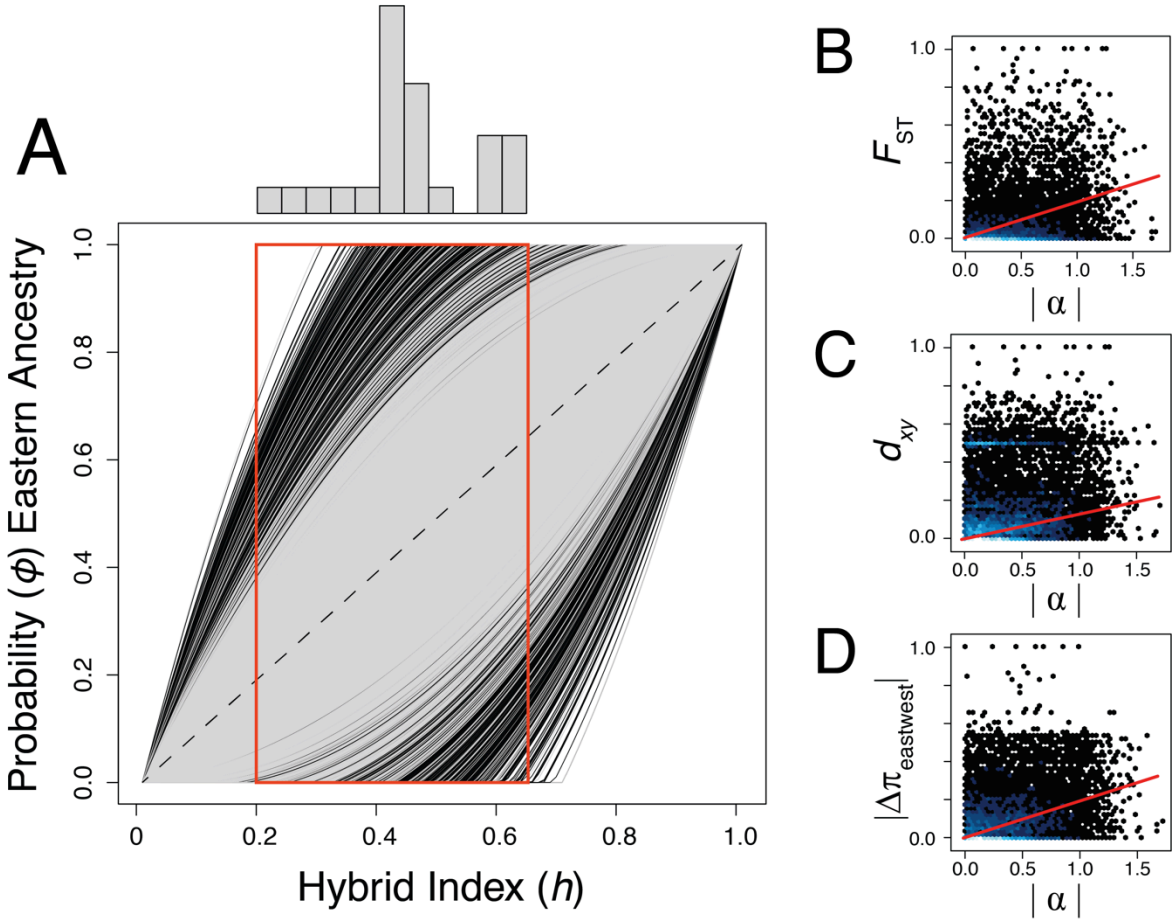


Figure 4. Introgression among genomic loci within hybrid *C. atrox* and depictions of the genomic relationship between allelic differentiation and introgression statistics. **A.** Results of genomic cline analysis in *bgc*, depicting the probability of western ancestry given background genomic introgression (i.e., hybrid index) for 7,031 SNP loci. The dashed line represents a perfect linear correlation between hybrid index and ancestry probability as expected under neutral introgression. Clines shown in black represent loci with excess ancestry from one parental population. The histogram above depicts relative frequencies of individual hybrid indices within the admixed population, and the red box is the range of these values on the genomic cline. **B.** Comparison of F_{ST} with the absolute value of the genomic cline center parameter, α. **C.** Comparison of d_{xy} and the absolute value of α. **D.** Comparison of $\Delta\pi_{eastwest}$ and α for all loci. Red lines in B-D represent statistically significant correlations between estimates of differentiation and introgression across the genome. Colors of dots in B-D reflect the density of loci, with lighter colors depicting greater quantities of loci.

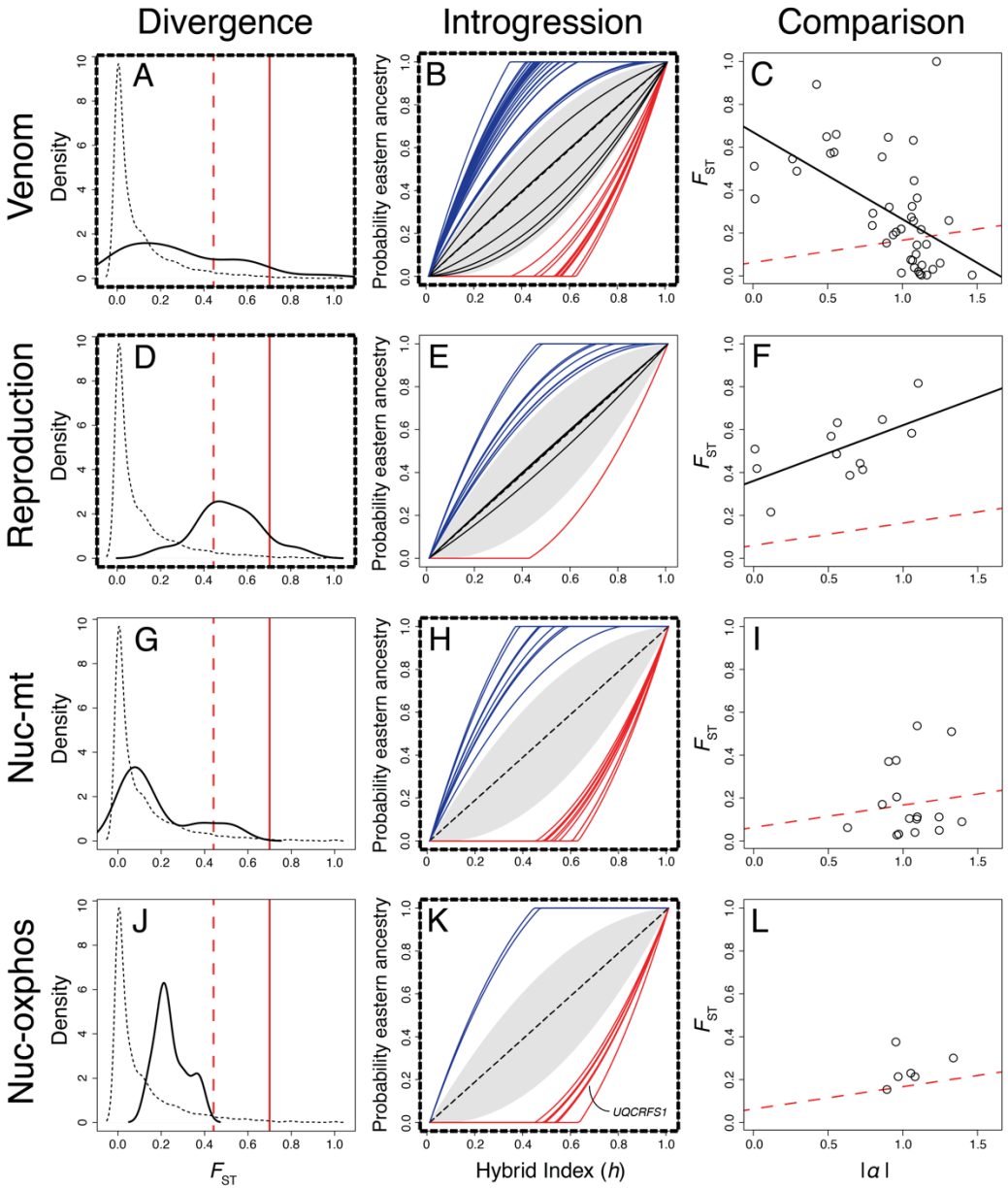


Figure 5. Divergence and introgression of enriched venom (A-C), reproduction (D-F), nuc-mt (G-I), and nuc-oxphos (J-L) candidate gene sets. Panels with bold, dashed borders represent outlier locus sets statistically enriched for specific candidate genes. **A, D, G, J.** Kernel-density plots of candidate gene (black line) and genome-wide (dashed line) F_{ST} estimates. Red dashed lines represent the 95th F_{ST} quantile and solid red lines represent the 99th F_{ST} quantile. **B, E, H, K.** Genomic clines of candidate genes with excess ancestry. The grey shaded region represents the genome background and is bounded by the positive and negative means of α. Bold colored lines represent outlier locus clines with evidence of excess eastern (blue) and western (red) ancestry. Black clines were candidate gene loci without evidence of excess ancestry. In panel K (nuc-oxphos genes), the locus-specific cline for *UQCRFS1* is labeled. **C, F, I, L.** Comparisons of locus-specific |α| and F_{ST} values for candidate gene sets. The trendlines in C and F depict statistically significant correlations. The red dashed lines depict the background genome correlation between F_{ST} and |α|, for comparison.

Tables

Table 1. Results of population genetic model comparison using the two-dimensional allele frequency spectrum (2D-AFS) between western and eastern *C. atrox* populations. The best-fit model and parameters are in bold. Visual comparison of the 2D-AFS for the data and the best-fit model is provided in Fig. 2.

Model	AIC	ΔAIC	RL	w_i	log-lik	params	theta	nu1	nu2	m12	m21	T1	T2
Divergence and Isolation, Secondary Contact with Symmetric Migration	**639.1**	**0.0**	**1**	**0.78**	**-314.6**	**5**	**93.8**	**1.01**	**3.54**	**0.189**	**m12**	**0.404**	**0.703**
Divergence and Isolation, Secondary Contact with Asymmetric Migration	642.4	3.3	0.192	0.15	-315.2	6	77.8	1.23	4.08	0.144	0.196	0.119	1.475
Divergence with Symmetric Migration	645.3	6.2	0.045	0.035	-318.6	4	122.4	0.74	2.77	0.184	m12	0.659	-
Divergence with Ancient Symmetrical Migration, Isolation	646.2	7.1	0.029	0.022	-318.1	5	77.6	1.17	4.15	0.310	m12	1.496	0.148
Divergence with Ancient Asymmetrical Migration, Isolation	647.6	8.5	0.014	0.011	-317.8	6	105.4	0.89	2.85	0.233	0.398	0.956	0.071
Divergence with Asymmetric Migration	662.6	23.5	0	0	-326.3	5	112.8	0.67	3.29	0.480	0.032	0.761	-
Divergence and Isolation	683.5	44.4	0	0	-338.7	3	131.4	0.76	2.88	-	-	0.467	-
Growth Model, No Divergence	4498.4	3859.3	0	0	-2246.2	3	115.7	-	-	-	-	-	-
Neutral model, No Divergence	4736.9	4097.8	0	0	-2368.4	0	232.5	-	-	-	-	-	-

Abbreviations are as follows: AIC, Akaike information criterion; RL, relative likelihood; w_i, Akaike Weight; params, number of parameters in model; theta, $4N_{ref}\mu L$; nu1, effective population size of Western group; nu2, effective population size of Eastern group; m12, migration rate from eastern population to western population; m21, migration rate from western population one to eastern population; T1, scaled time between population split and the present or T2, the scaled time of secondary contact or isolation interval.

Table 2. Results of lineage delimitation using BFD*, testing two competing models with *C. scutulatus* as outgroup.

Model	Lineages	ML	Rank	BF
No *atrox* divergence + *scutulatus*	2	-2,978	2	-
Two *atrox* lineages + *scutulatus*	**3**	**-2,810**	**1**	**-336**

Abbreviations are as follows: ML, Marginal Likelihood; BF, Bayes Factor.

Appendix

RADseq data generation

We extracted genomic DNA using phenol-chloroform isoamyl alcohol extractions, quantified purified DNA products using a Qubit flourometer (Life Technologies, Grand Island, NY, USA), and loaded DNA into restriction digestion reactions. DNA was digested using Sau3AI and SbfI restriction enzymes and digested products were purified using AMPure beads (Invitrogen, Calsbad, CA, USA). Custom Illumina adapters with dual indices and an 8bp unique molecular identifier (UMI) sequence were then ligated to digested fragments. We size selected for fragments between 575-655 bp using a Blue Pippin Prep (Sage Science, Beverly, MA, USA), targeting roughly 20,000 genomic loci. Size-selected libraries were PCR amplified using barcoded primers and Phusion high-fidelity polymerase enzyme (New England Biolabs, Ipswich, MA, USA), purified using AMPure beads, and quantified using an Agilent Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). Following quantification, libraries were pooled together in equimolar ratios and sequenced on an Illumina HiSeq 2500 using 100 bp paired-end reads.

Posterior predictive simulation of population divergence under neutrality

We performed posterior predictive simulations (PPS) of population divergence under a strictly neutral model of mutation and drift using our software *GppFst*. To conduct PPS analysis of genome-wide allelic differentiation, we used 7,031 nuclear SNPs to estimate divergence time and population parameters for the parental eastern and western populations ($\tau_{\text{east-west}}$, θ_{west}, θ_{east}, and $\theta_{\text{east-west}}$) via Markov Chain Monte Carlo (MCMC) sampling implemented in the program SNAPP (Bryant et al. 2012). The mean posterior estimates of population genetic parameters

used for *GppFst* simulations from SNAPP were as follows: $\theta_{east} = 0.217$, $\theta_{west} = 0.0592$, $\theta_{east\text{-}west} = 0.0615$, and $\tau_{east\text{-}west} = 0.00578$. We ran the MCMC chain for a total of 500,000 generations, sampling every 1,000 generations. We assessed posterior convergence and stationarity using Tracer (for all parameters ESS > 500; (Drummond and Rambaut 2007) and discarded the first 125,000 (25%) steps as burn-in, leaving total of 375 MCMC samples used to generate a PPS F_{ST} distribution. For each MCMC step, we then simulated 100 independent loci with a length of 190 base pairs under a JC69 model using the R package phybase (Liu and Yu 2010) with random sampling of individuals from the empirical distributions of locus coverage for east and west samples, respectively. We calculated F_{ST} and d_{xy} for a randomly sampled polymorphic site for each simulated locus to generate a theoretical distribution of values, which allowed us to compare empirical distributions of F_{ST} and d_{xy} to identify loci that are poorly explained by the neutral model of divergence. Within this framework, we effectively accounted for multiple sources of uncertainty that may influence our empirical distributions of allelic differentiation, including divergence times, differences in θ parameters for both parental populations, variation in coverage across loci and between populations, and SNP ascertainment.

Bayesian genomic clines analysis

We ran *bgc* (Gompert and Buerkle 2012) using the genotype uncertainty model using four chains of 70,000 generations each, discarding 40,000 generations as burn-in, and recording parameter estimates from every fifth MCMC generation. Default settings for bgc were used, assuming free recombination between loci, and we set the sequence error probability parameter to 0.001. Parameter estimates from four chains were combined after confirming convergence and stationarity of parameters from each MCMC run. These specifications were used also for

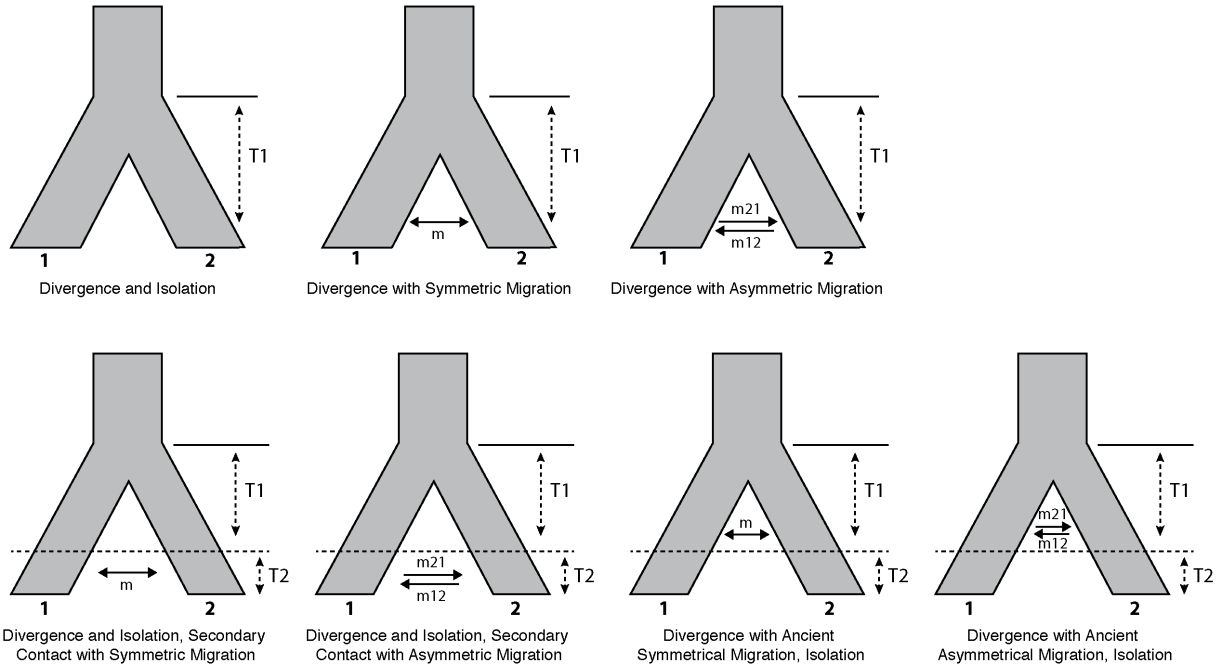
simulations of neutral introgression between randomly generated 'admixed' loci between eastern and western parental populations (see Methods).

Candidate gene set construction

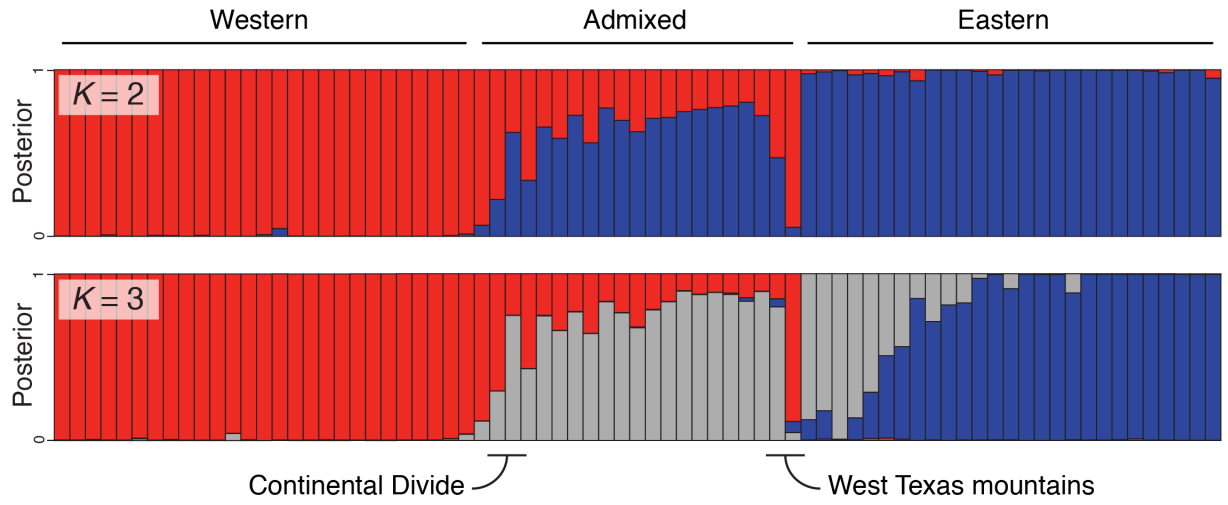
To identify regions with homology to known snake venom genes, we used sequences for 22 representative venom gene families from (Reyes-Velasco et al. 2015), to construct a blast database and searched all Cobra genome CDS sequences for all unique coding regions found up and downstream of variant regions, and performed a tblastx search to find putative venom gene family homologs, setting an e-value cutoff of 0.01. We used a similar methodology for coloration candidate genes, for which we obtained sequence for 28 genes involved in vertebrate chromatophores and color patterning (Hoekstra 2006; Hubbard et al. 2010; Irion et al. 2016). We curated a list of 'reproduction' candidate genes by combining gene sequences included in the 'regulation of oocyte development' and 'ovarian follicle development' gene ontology terms (528 total genes). We limited our search to these terms because they were the most closely related to the specific differences in reproductive output between populations of *C. atrox* (Spencer 2003). For the nuclear-encoded mitochondrial gene (nuc-mt) sequence set, we obtained all *Homo sapiens* sequences available in the MitoRes database (1063 genes;(Catalano et al. 2006)). Separately, we generated the more specific oxidative phosphorylation (oxphos) set using genes in the KEGG 'oxidative phosphrylation' pathway (120 genes; http://www.genome.jp/dbget-bin/get_linkdb?-t+genes+path:hsa00190), excluding genes encoded in the mitochondrial genome. We generated blast databases for each candidate sequence set, and performed tblastx searches against all up and downstream Cobra genes. For each search, we filtered results to retain a maximum of one target sequence per Cobra gene, and called putative homologs using the

e-value distribution, evaluating the point at which 1 – e-value plateaued, indicating consistent sequence similarity among all putative homologs to subject sequences. We tested for enrichment by determining if proportions of candidate genes were greater in outlier sets relative to all Cobra genes, and used Fisher's Exact Tests to determine if proportions differed significantly.

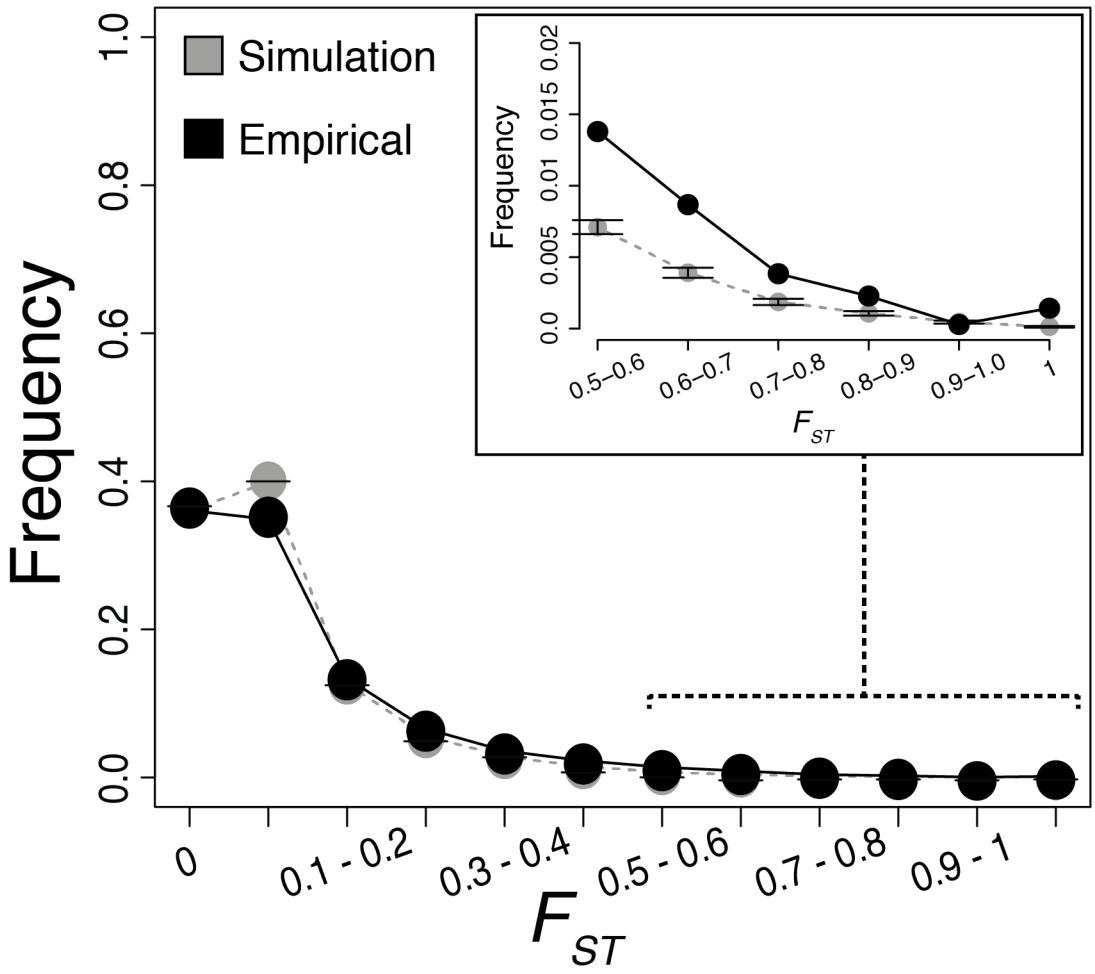
Supplementary Figures



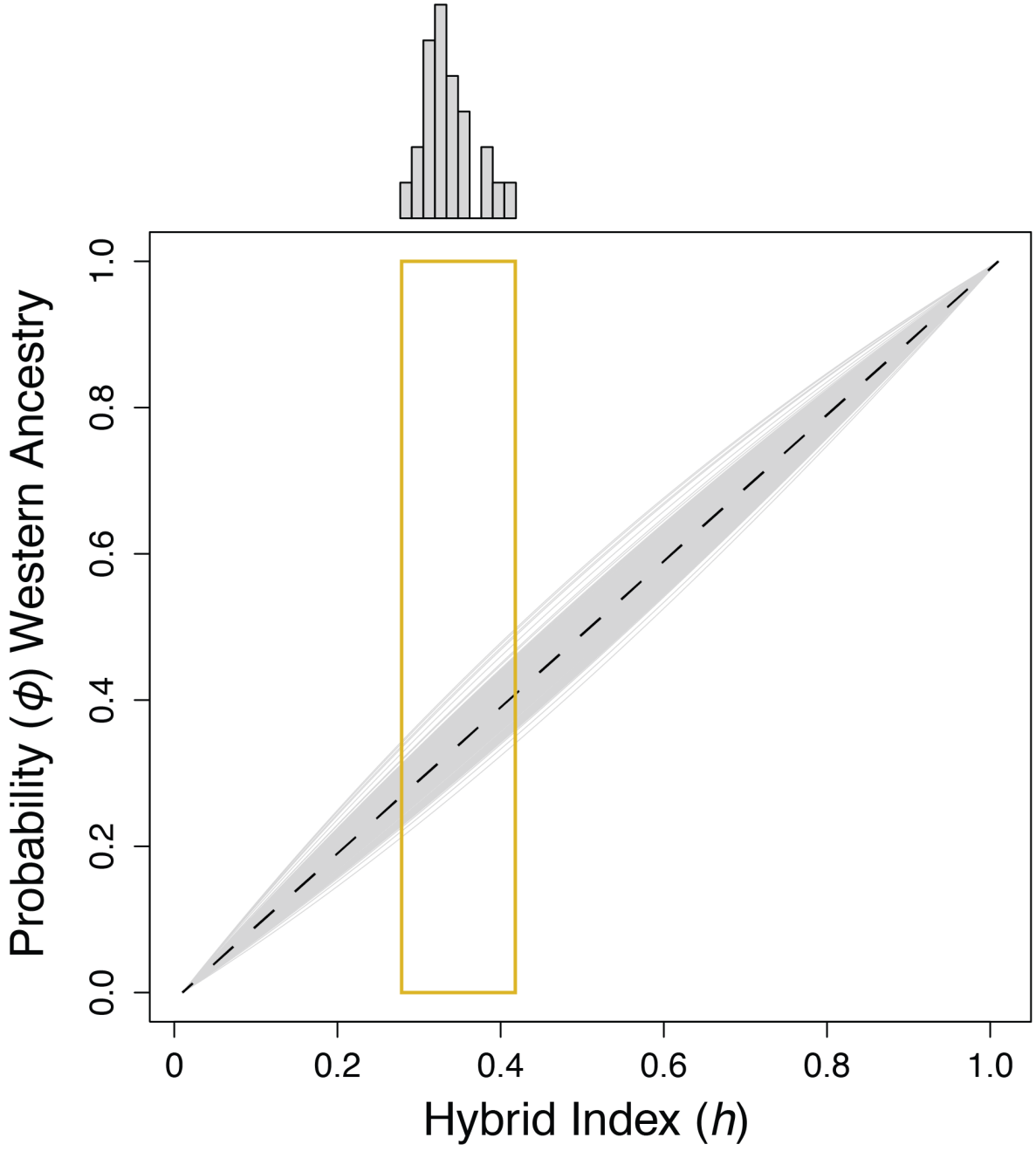
Supplementary Figure 1. Competing demographic models tested using the 2D allele frequency spectrum between eastern and western populations. Models involving no population divergence are not shown. See Table 1 and Figure 2 for detailed information about parameter estimates and the best-fit model.



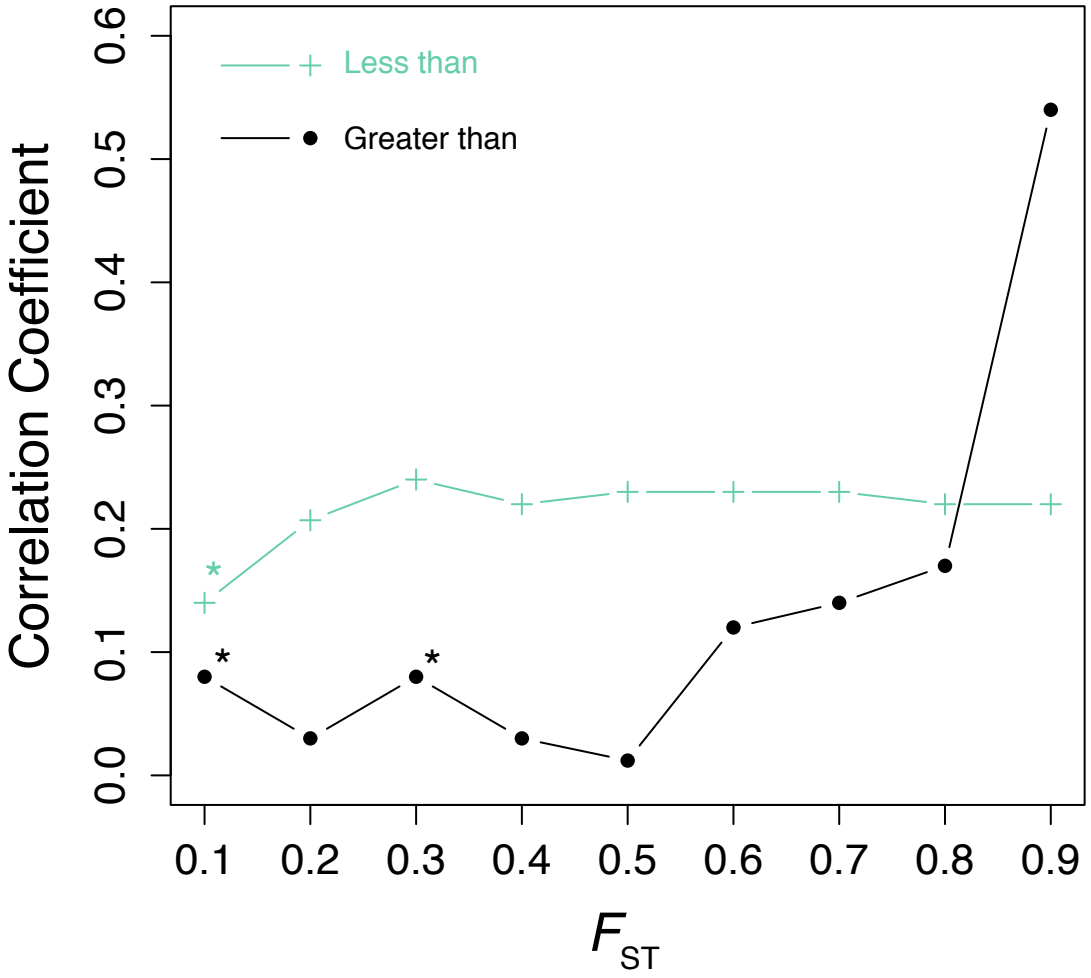
Supplementary Figure 2. Posterior probability assignments of each individual into genetic clusters inferred using STRUCTURE organized in a longitudinal gradient from west to east under $K = 2$ and $K = 3$ models. Geographic assignments used in divergence and introgression analyses are labeled below. The transitions across the Continental Divide and the mountains in western Texas are labeled under the individuals sampled at those localities.



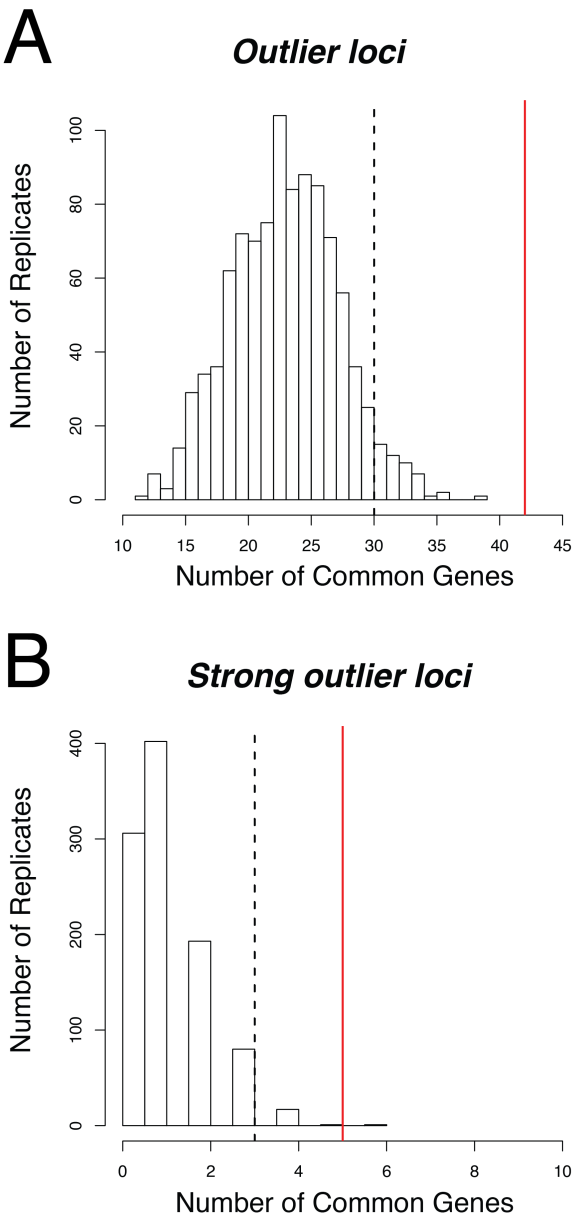
Supplementary Figure 3. Results of analyses of empirical and simulated distributions of F_{ST} in *GppFst*, showing proportions of loci from simulated (grey) and empirical (black) datasets that fall into bins of F_{ST} values in 0.1 intervals. The inset details proportions of loci with F_{ST} in bins greater than 0.5. Error bars on posterior predictive simulated points indicate the standard deviation of estimates across 100 replicates.



Supplementary Figure 4. Results of genomic cline analysis using an admixed population simulated from random alleles from each parental population for 7,031 loci. The dashed line represents a perfect linear correlation between hybrid index and ancestry probability as expected under neutral evolution. The histogram above depicts relative frequencies of individual hybrid indices within the admixed population, and the yellow box denotes the range of these values on the genomic cline.



Supplementary Figure 5. Correlation coefficients between F_{ST} and |α| for various intervals of F_{ST}. The black line/points indicate comparisons that were made including only loci equal to or greater than the value of the interval. The green line depicts comparisons where only loci with values less than the interval were used.



Supplementary Figure 6. Comparisons of observed numbers of overlapping genes and distributions of random samples of overlapping genes between genes linked to outlier loci (**A**) and strong outlier loci (**B**) from divergence and introgression analyses. Red lines denote the number of observed overlapping genes, and dashed black lines indicate the 95th percentile of the random sample distribution.

Supplementary Tables

Supplementary Table 1. Specimen data for samples used in this study. Where noted, samples were used previously in Schield et al. (2015).

Species	CA Number	Museum ID	Voucher	Country	State	County	Population	Reference
Crotalus atrox	CA0013	CAS 235728	RNF 2596	USA	CA	Imperial	West	This study
Crotalus atrox	CA0042		ENT A21	USA	AZ	Pinal	West	Schield et al. 2015
Crotalus atrox	CA0043		ENT A49	USA	AZ	Pinal	West	Schield et al. 2015
Crotalus atrox	CA0046		ENT 11	USA	AZ	Maricopa	West	Schield et al. 2015
Crotalus atrox	CA0048		ENT 7	USA	AZ	Maricopa	West	Schield et al. 2015
Crotalus atrox	CA0049	LACM 150957		USA	AZ	Pima	West	Schield et al. 2015
Crotalus atrox	CA0082	ROM 18144		USA	CA	Riverside	West	Schield et al. 2015
Crotalus atrox	CA0112	UTAR 50396	CLS 384	USA	AZ	Cochise	West	This study
Crotalus atrox	CA0113	UTAR 50402	CLS 388	USA	AZ	Cochise	West	This study
Crotalus atrox	CA0179	UTAR 50405	CLS 346	USA	AZ	Cochise	West	This study
Crotalus atrox	CA0237		CA039	USA	AZ	Pima	West	This study
Crotalus atrox	CA0260		CA062	USA	AZ	Pima	West	This study
Crotalus atrox	CA0272		CA074	USA	AZ	Pima	West	This study
Crotalus atrox	CA0274		CAPR 002	USA	AZ	Pima	West	This study
Crotalus atrox	CA0276		CAPR 004	USA	AZ	Pima	West	This study
Crotalus atrox	CA0278		CAPR 006	USA	AZ	Pima	West	This study
Crotalus atrox	CA0280		CAPR 008	USA	AZ	Pima	West	This study
Crotalus atrox	CA0281		CAPR 009	USA	AZ	Pima	West	This study
Crotalus atrox	CA0282		CAPR 010	USA	AZ	Pima	West	This study
Crotalus atrox	CA0294		CASZ-042	USA	AZ	Pima	West	This study
Crotalus atrox	CA0300		Cax001	USA	AZ	Cochise	West	This study
Crotalus atrox	CA0304		Cax005	USA	AZ	Cochise	West	This study
Crotalus atrox	CA0306		Cax007	USA	AZ	Santa Cruz	West	This study
Crotalus atrox	CA0307		Cax008	USA	AZ	Cochise	West	This study
Crotalus atrox	CA0308		Cax009	USA	AZ	Cochise	West	This study
Crotalus atrox	CA0335		DRS0018	USA	AZ	Yavapai	West	This study
Crotalus atrox	CA0345	NNTRC A29		USA	AZ	Cochise	West	This study
Crotalus atrox	CA0018		RWV 2001-09	USA	TX	Jeff Davis	Admixed	Schield et al. 2015
Crotalus atrox	CA0021		RWV 2001-13	USA	NM	Sierra	Admixed	Schield et al. 2015
Crotalus atrox	CA0022		RWV 2001-14	USA	NM	Dona Ana	Admixed	Schield et al. 2015
Crotalus atrox	CA0109	UTAR 50649	CLS 381	USA	NM	Hidalgo	Admixed	This study
Crotalus atrox	CA0110	UTAR 50445	CLS 382	USA	NM	Hidalgo	Admixed	Schield et al. 2015
Crotalus atrox	CA0111	UTAR 50650	CLS 383	USA	NM	Hidalgo	Admixed	Schield et al. 2015
Crotalus atrox	CA0114	UTAR 50418	CLS 393	USA	NM	Hidalgo	Admixed	This study
Crotalus atrox	CA0116	UTAR 50385	CLS 413	USA	AZ	Cochise	Admixed	Schield et al. 2015
Crotalus atrox	CA0117	UTAR 50652	CLS 414	USA	AZ	Cochise	Admixed	This study

Crotalus atrox	CA0120	UTAR 50407	CLS 419	USA	AZ	Cochise	Admixed	This study
Crotalus atrox	CA0130	UTAR 50409	CLS 240	USA	NM	Hidalgo	Admixed	Schield et al. 2015
Crotalus atrox	CA0132	UTAR 50411	CLS 244	USA	NM	Hidalgo	Admixed	Schield et al. 2015
Crotalus atrox	CA0133	UTAR 50412	CLS 247	USA	NM	Hidalgo	Admixed	This study
Crotalus atrox	CA0139	UTAR 50398	CLS 264	USA	AZ	Cochise	Admixed	Schield et al. 2015
Crotalus atrox	CA0146	UTAR 50399	CLS 282	USA	AZ	Cochise	Admixed	Schield et al. 2015
Crotalus atrox	CA0150	UTAR 50422	CLS 287	USA	NM	Hidalgo	Admixed	This study
Crotalus atrox	CA0151	UTAR 50376	CLS 290	USA	AZ	Cochise	Admixed	Schield et al. 2015
Crotalus atrox	CA0158	UTAR 50428	CLS 298	USA	NM	Hidalgo	Admixed	Schield et al. 2015
Crotalus atrox	CA0160	UTAR 50429	CLS 300	USA	NM	Hidalgo	Admixed	Schield et al. 2015
Crotalus atrox	CA0161	UTAR 50426	CLS 301	USA	NM	Hidalgo	Admixed	This study
Crotalus atrox	CA0171	UTAR 50404	CLS 344	USA	AZ	Cochise	Admixed	Schield et al. 2015
Crotalus atrox	CA0339		DRS0016	USA	TX	Reeves	Admixed	This study
Crotalus atrox	CA0028		RWV 2001-22	USA	TX	Jeff Davis	East	Schield et al. 2015
Crotalus atrox	CA0039		BLC 27	USA	NM	Sierra	East	Schield et al. 2015
Crotalus atrox	CA0063		TJL601	USA	TX	Llano	East	Schield et al. 2015
Crotalus atrox	CA0065		TJL868	USA	TX	Potter	East	Schield et al. 2015
Crotalus atrox	CA0068		TJL347	USA	TX	Val Verde	East	This study
Crotalus atrox	CA0069		TJL348	USA	TX	Val Verde	East	Schield et al. 2015
Crotalus atrox	CA0073		JJ	USA	TX	Culberson	East	Schield et al. 2015
Crotalus atrox	CA0096		RLG381	USA	TX	LaSalle	East	Schield et al. 2015
Crotalus atrox	CA0097		RLG367	USA	TX	Val Verde	East	Schield et al. 2015
Crotalus atrox	CA0098		RLG380	USA	TX	LaSalle	East	Schield et al. 2015
Crotalus atrox	CA0099		RLG390	USA	TX	Dimmit	East	Schield et al. 2015
Crotalus atrox	CA0100		RLG404	USA	TX	Zavala	East	Schield et al. 2015
Crotalus atrox	CA0193		DRS0002	USA	TX	Shackelford	East	Schield et al. 2015
Crotalus atrox	CA0194		DRS0003	USA	TX	Palo Pinto	East	Schield et al. 2015
Crotalus atrox	CA0195		DRS0005	USA	TX	Parker	East	Schield et al. 2015
Crotalus atrox	CA0196		DRS0007	USA	TX	Tom Green	East	Schield et al. 2015
Crotalus atrox	CA0342	NNTRC A26		USA	TX	Nueces	East	This study
Crotalus atrox	CA0343	NNTRC A21		USA	TX	Kleburg	East	This study
Crotalus atrox	CA0344	NNTRC A51		USA	TX	Kleburg	East	This study
Crotalus scutulatus	CS0121		Csx001	USA	AZ	Graham		This study
Crotalus scutulatus	CS0125		Csx004	USA	AZ	Cochise		This study
Crotalus scutulatus	CS0126		Csx005a	USA	AZ	Pima		This study
Crotalus scutulatus	CS0130	CAS 228094		USA	CA	San Bernardino		This study
Crotalus scutulatus	CS0131		CLS 800	USA	CA	San Bernardino		This study
Crotalus scutulatus	CS0132		CLS 811	USA	CA	Los Angeles		This study

Supplementary Table 2. Numbers of putative homologs to venom, reproduction, coloration, nuc-mt, and nuc-oxphos candidate gene sets from all orthologous Cobra genome regions, as well as specific outlier locus sets. Proportions of candidate genes in each outlier locus set and results of Fisher's Exact tests are provided below. Outlier sets in bold were significantly enriched.

Set	All genes	Venom genes	Reproduction genes	Coloration genes	Nuc-mt genes	Nuc-oxphos genes
All loci	5992	265	206	25	199	56
Divergence	190	18	14	0	11	2
Strong divergence	38	4	2	0	2	0
Introgression	312	32	12	1	19	8
Strong introgression	66	4	3	0	8	1

Fisher's Exact Test - Venom

Set	Prop. venom	p-value
All loci	0.044	-
Divergence	**0.095**	**0.005071**
Strong divergence	0.110	0.1028
Introgression	**0.103**	**7.62E-05**
Strong introgression	0.061	0.5418

Fisher's Exact Test - Reproduction

Set	Prop. reproduction	p-value
All loci	0.034	-
Divergence	**0.074**	**0.0163**
Strong divergence	0.053	0.3874
Introgression	0.038	0.6357
Strong introgression	0.045	0.501

Fisher's Exact Test - Coloration

Set	Prop. coloration	p-value
All loci	0.004	-
Divergence	0	1
Strong divergence	0	1
Introgression	0.003	1
Strong introgression	0	1

Fisher's Exact Test - Nuc-mt

Set	Prop. nuc-mt	p-value
All loci	0.033	-
Divergence	0.058	0.1019
Strong divergence	0.053	0.3717
Introgression	**0.061**	**0.01802**
Strong introgression	**0.121**	**0.009745**

Fisher's Exact Test - Nuc-oxphos

Set	Prop. nuc-oxphos	pval
All loci	0.009	-
Divergence	0.011	0.6992
Strong divergence	0	1
Introgression	**0.026**	**0.01414**
Strong introgression	0.015	0.4679

Chapter 4

Cryptic genetic diversity, population structure, and gene flow in the Mojave rattlesnake (*Crotalus scutulatus*)

Drew R. Schield[a], Daren C. Card[a], Richard H. Adams[a], Andrew B. Corbin[a], Tereza Jezkova[b], Nicole R. Hales[a], Jesse M. Meik[c], Blair W. Perry[1], Carol L. Spencer[d], Lydia L. Smith[d], Gustavo Campillo García[e], Nassima M. Bouzid[d,f], Jason L. Strickland[g], Christopher L. Parkinson[g], Miguel Borja-Jiménez[h], Camaliel Castañeda-Gaytán[h], Robert W. Bryson[f], Oscar A. Flores-Villela[e], Stephen P. Mackessy[i], and Todd A. Castoe[a,]*

[a]Department of Biology, 501 S. Nedderman Drive, University of Texas at Arlington, Arlington, TX 76019 USA

[b]Department of Ecology and Evolutionary Biology, 1041 E. Lowell Street, University of Arizona, Tuscon, AZ, 85721 USA

[c]Department of Biological Sciences, Tarleton State University, 1333 W. Washington Street, Stephenville, TX, 76402 USA

[d]Museum of Vertebrate Zoology, 3101 Valley Life Sciences Building, University of California, Berkeley, CA 94720 USA

[e]Museo de Zoología, Department of Evolutionary Biology, Faculta de Ciencias, Universidad Nacional Autónoma de México, External Circuit of Ciudad Universitaria, Mexico City, Mexico

[f]Department of Biology and Burke Museum of Natural History and Culture, University of Washington, Box 351800, Seattle, WA 98195 USA

[g]Department of Biology, Biological Sciences Building, 4110 Libra Drive, University of Central Florida, Orlando, FL 32816 USA

[h]Facultad de Ciencias Biológicas, Universidad Juárez del Estrado de Durango, Gómez Palacio, Durango, Mexico

[i]School of Biological Sciences, 501 20th Street, University of Northern Colorado, Greeley, CO 80639 USA

Abstract

The Mojave rattlesnake (*Crotalus scutulatus*) inhabits deserts and arid grasslands of the western United States and Mexico. Despite considerable interest in its highly toxic venom and the recognition of two subspecies, no molecular studies have characterized range-wide genetic diversity and population structure or tested species limits within *C. scutulatus*. We used mitochondrial DNA and thousands of nuclear loci from double-digest restriction site associated DNA sequencing to infer population genetic structure throughout the range of *C. scutulatus*, and to evaluate gene flow between structured populations. We find strong support for several divergent mitochondrial and nuclear clades of *C. scutulatus*, including splits coincident with two major phylogeographic barriers: the Continental Divide and the uplift of the Mexican Plateau. We apply Bayesian clustering, phylogenetic inference, and coalescent-based species delimitation to our nuclear genetic data to test mitochondrial hypotheses of population structure and monophyly of lineages, and we performed demographic analyses to infer patterns of genetic structure and gene flow. Collectively, our data provide strong support for previously undocumented diversity within *C. scutulatus*, and genetically defined populations and previously defined subspecies ranges do not correspond. Finally, we use approximate Bayesian computation to test hypotheses of divergence among multiple rattlesnake species groups distributed across the Continental Divide, and find evidence for co-divergence at this boundary during the mid-Pleistocene.

Introduction

The Mojave Rattlesnake (*Crotalus scutulatus*) is widely distributed throughout the Mojave, Sonoran, and Chihuahuan deserts and associated grasslands of the western United States and

Mexico (Klauber 1956). It inhabits primarily dry scrubland habitat in flat terrain but also occurs in montane woodlands near the southern end of its distribution (Campbell and Lamar 2004), and occupies elevations from sea level to over 2,100 meters. *Crotalus scutulatus* is considered among the most dangerous North American pitvipers due to its aggressive disposition, high venom yield, and highly neurotoxic venom (Cate and Bieber 1978; Hardy 1983; Sanchez et al. 2005; Mackessy 2008). For these reasons, and with the intention of generating improved treatments for snakebite, *C. scutulatus* has been the subject of numerous venom studies (e.g., (Glenn and Straight 1977; Glenn et al. 1983; Massey et al. 2012; Durban et al. 2013; Smith and Mackessy 2016). These studies have largely sought to characterize populations with and without expression of neurotoxic venom compounds. Despite efforts to understand the toxicity and geographic variation of *C. scutulatus* venom, remarkably little is known regarding the overall genetic structure and variation throughout the range of this species.

C. scutulatus has been included in several broad phylogenetic analyses of pitvipers, and has been well supported as the sister lineage to the Western Rattlesnake (*Crotalus viridis* + *oreganus*) species complex (Murphy et al. 2002; Castoe and Parkinson 2006; Reyes-Velasco et al. 2013), from which it diverged during the Pliocene (Reyes-Velasco et al. 2013). Two phenotypically distinctive subspecies of *C. scutulatus* are currently recognized: the Mojave Rattlesnake (*C. s. scutulatus*) and the Huamantlan Rattlesnake (*C. s. salvini*), (Campbell and Lamar 2004). The majority of the range of *C. scutulatus* is allocated to the northern subspecies *C. s. scutulatus,* which ranges from southern California to central Mexico. The southern subspecies *C. s. salvini* has a comparatively smaller range in the south-central Mexican states of Hidalgo, Tlaxcala, Estado de Mexico, Puebla, and Veracruz. Although Klauber (1956) studied intraspecific

variation in *C. scutulatus* morphology, no studies have assessed genetic relationships among populations and subspecies. Considerable cryptic diversity within the species is plausible given the large geographic range and distinctive ecoregions inhabited by *C. scutulatus*. Given the large range, medical relevance, and absence of any molecular studies of intraspecific variation of *C. scutulatus*, a range-wide molecular appraisal of the species will serve as a foundation for broad research and medical interests.

Here we evaluate patterns of genetic diversity and structure of *C. scutulatus* throughout its range and provide the first molecular-based inferences of *C. scutulatus* phylogeography and population genetics using a combined dataset of a mitochondrial gene and thousands of genome-wide SNPs. We test the hypothesis that *C. s. scutulatus* and *C. s. salvini* are distinct evolutionary lineages and explore potential cryptic diversity within *C. scutulatus* that is not recognized by subspecies taxonomy. We also estimate population structure and gene flow between populations and characterize the biogeographic and demographic patterns within lineages of *C. scutulatus*. More broadly, we test the hypothesis that the Continental Divide represents a partial barrier to gene flow between subpopulations of *C. s. scutulatus*, as has been observed in the closely related Western Diamondback Rattlesnake (*C. atrox*; (Castoe et al. 2007; Schield et al. 2015) and other snake taxa (Myers et al. 2016). *Crotalus scutulatus* also traverses a breadth of topographies and ecoregions in Mexico, including the uplift of the Mexican Plateau, and we also tested for divergence and gene flow across this known biogeographic barrier. Overall our results highlight substantial and previously unappreciated genetic structure and lineage diversity within *C. scutulatus* that is valuable for interpreting data on venom variation, snakebite treatment, and ultimately will help guide future taxonomic revision of this group.

Materials and Methods

Taxon sampling and DNA extraction

We obtained tissue samples from 141 *Crotalus scutulatus* specimens, including representatives of both recognized subspecies (Fig. 1; Supplementary Table 1). We sampled specimens from diverse localities to maximize our ability to estimate population structure and genetic diversity range-wide. Skin or liver tissue was dissected and snap frozen in liquid nitrogen, or stored in DNA lysis buffer or in ethanol. Non-lethal caudal punctures were used to obtain blood tissue from some specimens, and blood was snap-frozen in liquid nitrogen. We isolated genomic DNA from blood and whole tissue using phenol-chloroform-isoamyl alcohol and NaCl-isopropanol precipitation extractions (Miller et al. 1988).

Mitochondrial and nuclear DNA sequence generation

We used PCR to amplify a fragment of the mitochondrial NADH4 (ND4) gene and downstream tRNAs from all samples, using the primers ND4 and Leu (Arevalo et al. 1994). PCR products were purified using Serapure beads and were quantified using gel electrophoresis and a Qubit Fluorometer (Life Technologies, Grand Island, NY, USA). Purified PCR products were sequenced in both directions using amplification primers and BigDye on an ABI 3730 capillary sequencer (Life Technologies), and ND4 sequences for outgroup taxa were retrieved from our previous studies (Castoe et al. 2007; Schield et al. 2015) or from GenBank (see Supplementary Online Table 1 for accession numbers).

We chose a subset of *C. scutulatus* sequenced for mitochondrial markers to generate double-digest RADseq libraries (n = 48), using the Peterson et al. (2012) protocol with minor

adjustments detailed in (Schield et al. 2015). In brief, we began with 0.5 µg genomic DNA per sample and performed digestions overnight at 37°C using rare-cutting (*SbfI*) and common-cutting (*Sau3AI*) restriction enzymes. Digested samples were purified with Ampure (Invitrogen) beads; purified products were eluted in 40 µL of TE and quantified using a Qubit Fluorometer. Based on these quantifications, we organized samples by concentration, divided samples into groups of six to eight, and standardized input DNA for ligation reactions based on the lowest concentration among within-group samples (these were variable). We then ligated double-stranded indexed DNA adapters (including 8 bp unique molecular identifier regions; UMIs), and pooled samples into their respective groups, with each sample having a specific barcoded adapter. Pooled samples were purified and size-selected using a 1.5% agarose cassette on a Blue Pippin Prep (Sage Science) for fragments within a range of 575 – 655 bp. Based on *in silico* digestion and size selection using the Burmese Python genome (Castoe et al. 2013), we estimated this protocol would target ~20,000 loci. Size selected samples were amplified using indexed primers to provide a second index specific to each pooled sample. Indexed pools were mixed in equimolar ratios and were sequenced together using 100 bp single-end reads on an Illumina HiSeq 2500.

Estimates of mitochondrial gene diversity, phylogeny, and demography

Raw mitochondrial ND4 sequence chromatograms were edited and consensus sequences were generated using Geneious v6.1.6 (Biomatters Ltd., Aukland, NZ). Sequences were aligned using MUSCLE v3.8 (Edgar 2004) and trimmed manually to remove spurious and potentially erroneous low quality base calls. The final ND4 alignment consisted of 808 bases for all individuals, including outgroup taxa, with no indels. We estimated mitochondrial relationships

within *C. scutulatus* using MrBayes v3.2.1 (Huelsenbeck and Ronquist 2001), after inferring best-fit models of evolution for partitioned 1st, 2nd, and 3rd codon positions using the Bayesian Information Criterion and the Greedy algorithm implemented in PartitionFinder v1.1.1 (Lanfear et al. 2012). The best-fit partitioning scheme was an independent partition for each of the three codon positions (HKY + Γ for 1st and 3rd codon positions, F81 + Γ+I for 2nd codon positions); this partitioned model was used in all MrBayes runs. Two initial trial runs were performed using our total dataset, including 141 *C. scutulatus* individuals and four outgroup taxa to identify unique haplotypes. We then condensed the alignment to single representatives of each haplotype and performed four independent runs of MrBayes on this condensed dataset, each consisting of 10^7 generations with samples taken every 500 generations and four MCMC chains (one cold and three heated). We discarded the first 25% of MCMC samples in each run as burnin, as the potential scale reduction factor (PSRF) values, parameter estimates, and marginal likelihood estimates indicated that runs had converged by this period, which we assessed using Tracer v1.6 (Drummond and Rambaut 2007). Effective sample sizes exceeded 2,000 for all parameters indicating that the posterior probability distribution had been effectively sampled in each run. We generated a 50% majority rule tree using combined tree estimates from post-burnin samples.

We estimated parameters of the Isolation-Migration (IM) model using IMa2 (Hey 2010) to infer the divergence time, migration rates, and effective population sizes using the complete mtDNA alignment. Here, we refer to mitochondrial clades and 'populations' interchangeably; spatial organization of mitochondrial clades was consistent with populations in distinct geographic regions (see Results). The population structure and guide tree used in IM analyses mirrored the mitochondrial gene tree (see Fig. 1) and included four populations: i) a northern population

consisting of samples from California, Arizona, and New Mexico ('Mojave-Sonoran' hereafter), ii) a 'Chihuahuan' population including samples from Texas, Chihuahua, Coahuila, and northern Durango, iii) a population occupying higher elevation regions of the central Mexican Plateau ('Central Plateau' hereafter), and iv) a population from Puebla, Tlaxcala, and Veracruz, which represents the currently recognized range of *C. s. salvini*. For simplicity, we refer to this population as *salvini*. We used the HKY model of nucleotide substitution, a mutation rate scalar of 1.4×10^{-6} mutations per year (Castoe et al. 2007), and a generation time of 3 years for parameter estimation and conversions to demographic units. We performed four independent runs of IMa2 on these data, setting a 7.5×10^{6} generation burnin followed by 1.5×10^{7} sampled generations, from which we sampled the posterior parameter estimates every 100 generations. We assessed proper mixing based on ESS values that exceeded 1,000 for all parameters and determined convergence of posterior distributions by comparing independent runs.

We estimated mitochondrial genetic diversity directly using the average within-group pairwise distance, nucleotide diversity (π), Watterson's estimator ($\hat{\theta}$), and haplotype diversity calculated in DNAsp v5 (Librado and Rozas 2009). We also tested for patterns of allelic diversity in each population consistent with population expansion using Fu's *Fs* (Fu 1997). We obtained p-values for *Fs* by performing 1,000 coalescent simulations and computing the probability of observing values equal to or less than each simulated value.

Nuclear analyses of population structure, phylogeny, and demography

We processed raw Illumina sequencing reads first using the Stacks v1.35 clone_filter module (Catchen et al. 2013) to remove PCR clones using the UMI regions included in our sequence library adapters. We then trimmed the 8 bp UMIs from reads using the Fastx-Toolkit trimmer

(Hannon 2014). Trimmed reads were demultiplexed into individual samples using the Stacks process_RADtags function, which also trimmed the 6 bp barcode region of each read. We performed *de novo* assembly and variant calling using the Stacks ustacks, cstacks, and sstacks modules, and used the downstream rxstacks filter module to retain only biologically plausible loci. We then used the populations module to calculate heterozygosity and private alleles and generate sequence alignments by parsing the dataset to the individual level. We also performed pairwise analyses of allelic differentiation (F_{ST}) between populations to evaluate population structure. We filtered data to retain only loci that had at least 5x read depth per individual and for which data were present in at least 30 samples (Stacks options 'm' = 5 and 'p' = 30).

We used STRUCTURE (Pritchard et al. 2000) to estimate broad scale population structure and admixture within our nuclear SNP dataset. Initial runs to determine a reduced range of likely models to explain the data were performed using a *K* range of 1-10 clusters, and we performed 3 replicate runs per model of *K* with a burnin of 10,000 and 10,000 sampled MCMC generations. The results of preliminary runs were processed using Structure Harvester (Earl and Vonholdt 2012), from which we determined the best-fit model to be $K = 3$ using the Evanno method (Evanno et al. 2005). We then performed longer triplicate runs across a smaller range of $K = 2$-5, with a 10,000 generation burnin and 100,000 sampled generations, and assessed model likelihoods in Structure Harvester. To complement STRUCTURE analyses and to investigate evidence of genetic structure without model assumptions, we performed Discriminant Analysis of Principal Components (DAPC; (Jombart et al. 2010)). We first performed *K*-means clustering of the dataset specifying a possible 30 genetic clusters and estimated likelihoods over 1,000,000 iterations. We then evaluated model likelihoods using Bayesian Information Criterion (BIC), as

recommended by Zhao (2006) and (Lee et al. 2009) to determine the best-supported number of clusters.

We estimated phylogenetic relationships using Bayesian MCMC inference on a concatenated nuclear dataset in BEAST v2.1.3 (Bouckaert et al. 2014) with a GTR + Γ substitution model. We ran two independent BEAST runs for a total of 50 x 10^6 generations each, and discarded the first 25% (12.5 x 10^6 generations) as burn-in based on likelihood stationarity and parameter effective sample sizes (ESS) analyzed in Tracer v1.6 (Drummond and Rambaut 2007). We combined the two runs into a single posterior distribution consisting of 75 x 10^6 MCMC samples and generated a maximum clade credibility tree using TreeAnnotater (Bouckaert et al. 2014) based on mean heights, which is shown in Fig. 1B.

To assess evidence for reproductive isolation across *C. scutulatus* populations and to test mtDNA and nuclear tree-based hypotheses of either four or three major clades within *C. scutulatus*, we conducted Bayesian species delimitation using bpp (Yang and Rannala 2010) with 922 randomly sampled loci from our RADseq dataset (to make analyses computationally tractable). We set the prior distributions for the tree height parameter (τ) and effective population size parameters (θ) based on our mtDNA phylogeny and the average pairwise genetic distance (π) measured across our entire RADseq dataset. We set the shape and scale parameters of the gamma prior distribution on θ to 1.2 and 1.0, respectively, which reflects an expected θ of 0.12 (average π = 0.12). For the prior on τ, we set the shape and scale parameters to 1.6 and 10000, respectively, to reflect the tree height for these four clades that was estimated from our mtDNA phylogeny (distance to ancestral node = 0.00016). We also set the starting topology to match the mtDNA topology ((Mojave-Sonoran, Chihuahuan), (Central Plateau, *salvini*)) and conducted

unguided species delimitation using prior algorithm0. We ran the MCMC chain for 110,000 iterations, sampling every 5th iteration, and discarded the first 10,000 iterations as burnin.

We tested for evidence of gene flow between lineages of *C. scutulatus* using Patterson's D statistics (Durand et al. 2011), which quantify shared derived alleles that are expected to arrive via introgression rather than incomplete lineage sorting. We performed two analyses that tested for introgression across two biogeographic barriers: the Continental Divide and the uplift of the Mexican Plateau. For both analyses, we used the four-population 'CalcD' algorithm implemented in the R package evobiR (https://cran.r-project.org/web/packages/evobiR/index.html; see (Streicher et al. 2014) for additional details). The 'Continental Divide' analysis included Central Plateau *C. scutulatus* as the outgroup and the Chihuahuan group as the first ingroup. We then split the Mojave-Sonoran group into two subclades consisting of samples from New Mexico and eastern Arizona ('Mojave-Sonoran B') and samples from western Arizona and California ('Mojave-Sonoran A'; Fig. 2). We generated consensus sequences for each population using Stacks, specifying a minimum read depth of five per locus per individual ('-m' = 5), requiring each population to contain data for a given SNP locus ('-p' = 4), and requiring each SNP to be represented by at least 75% of individuals within a population ('-r' = 0.75). These parameters resulted in 6,058 SNPs for the 'Continental Divide' analysis. The 'Mexican Plateau' analysis required an outgroup outside of *C. scutulatus* to test for gene flow between the Central Plateau and Chihuahuan populations, so we generated RADseq data from four individual *C. viridis* (Supplementary Table 1), the sister lineage to *C. scutulatus*, following the same library preparation and sequencing protocols. The Stacks parameters detailed

above yielded 5,077 SNPs for the 'Mexican Plateau' analysis. Schematic representations and details of each analysis are provided in Fig. 2.

Inferences of historical biogeography and population expansion

We estimated the distribution of *C. scutulatus* at the height the Last Glacial Maximum (LGM) during the Pleistocene. We were specifically interested to determine if patterns of genetic divergence and structure observed in our datasets were consistent with predictions of historical geographic isolation during Pleistocene glacial conditions (represented by LGM conditions). We obtained geographic coordinates from museum records for 740 *C. scutulatus* specimens from Vertnet (www.vertnet.org) and combined these with coordinates from our genetic dataset. We then used ecological niche modeling (ENM) to estimate the present-day range of *C. scutulatus*, as implemented in MAXENT v3.3.3k (Phillips et al. 2006), which extracts environmental data associated with occurrence records and predicts the suitability of climatic conditions within a given range using prediction algorithms (Elith et al. 2006).

Because occurrence records were biased towards heavier sampling in the United States than in Mexico, we subsampled occurrence records to account for potential biases in ENM analysis, setting a minimum distance between samples of 0.1°, 0.2°, 0.3°, and 0.4°, then determined if subsampling schemes quantitatively altered ENM results. The environmental data were represented by temperature and precipitation variables from the WorldClim dataset v1.4 with a 2.5-minute resolution (Hijmans et al. 2005). Following the methodology of Jezkova et al. (2011), we removed 8 highly correlated variables (*i.e.*, correlation coefficient > 0.9), resulting in the selection of 11 climatic variables, and we confined models to the southwestern region of North America. We used default MAXENT parameters and used cross-validation as a replicated run

type. We ran 20 replicates for each model, determined an average model using logistic probability classes of climatic niche suitability, and visualized this model in ArcGIS v9.2 (ESRI). We used the receiver-operating characteristic (ROC) to determine an area under the curve (AUC) value to evaluate model performance, which range from 0.5 (i.e., random prediction) to 1 (perfect prediction; (Raes and ter Steege 2007). We then projected present-day models onto reconstructions of the LGM, with the assumption that the climatic niche of *C. scutulatus* has not changed between the LGM and present day (Elith et al. 2010). We used three ocean-atmosphere simulation models for environmental layers representing the LGM: CCSM4, MIROC-ESM, and MPI-ESM-P (www.worldclim.org).

Population genetic theory predicts that with range expansion, populations at or near to the expanding range-front will harbor lower genetic diversity (i.e., heterozygosity) than the ancestral population (Mayr 1942; Slatkin and Excoffier 2012), and we have observed this pattern studying RADseq loci in several empirical systems (Jezkova et al. 2015; Schield et al. 2015; Streicher et al. 2016). We tested the hypothesis that populations of *C. scutulatus* have experienced range expansion out of an ancestral region using linear models implemented in R (R Core Team 2017). Specifically, we evaluated whether estimates of individual heterozygosity and private alleles correlated with latitude and longitude.

Divergence dating and tests of co-divergence

We analyzed an expanding sample of *Crotalus* lineages using BEAST 2 (Bouckaert et al. 2014) to estimate divergence times for major splits between *C. scutulatus* lineages, and to test for evidence of co-divergence (i.e., correlated divergence events among taxa) of two or more distinct lineages across the *C. scutulatus* + *atrox* phylogeny, particularly between populations adjacent to

the Continental Divide. The expanded dataset used for these analyses included all *C. scutulatus* ND4 sequences (condensed to unique haplotypes), as well as ND4 sequences of western diamondback rattlesnakes (*C. atrox*), and several outgroup taxa (*C. ruber*, *C. horridus*, *C. molossus*, *Agkistrodon contortrix*; Supplementary Table 1). Sequences were aligned using MUSCLE and bases were trimmed at each end of the alignment to reduce missing data; the final alignment consisted of 807 bases. We estimated substitution models and partition assignments using PartitionFinder v1.1.1 (Lanfear et al. 2012), and partitioned the final alignment independently for each codon position. We applied an HKY model of sequence evolution to first and second codon position partitions and a TN93 model for the thirds codon position. To calibrate the tree we specified four clade constraints based on previous divergence estimates for pitvipers. The ancestor of *Agkistrodon*, the most distant outgroup used in this analysis, was constrained at 6 MYA, with a specified mean = 0.01 and standard deviation (SD) = 0.42. The ancestor of *Sistrurus*, the more closely related outgroup to *Crotalus*, was constrained with an offset of 8 MYA, a mean = 0.01, and SD = 0.76. These priors were specified using lognormal prior distributions and constraints according to Reyes-Velasco et al. (2013). We constrained the ancestral node for *C. atrox* + *C. ruber* with an offset of 3.2 MYA following Castoe et al. (2007), using a normal distribution with mean = 0 and SD = 1. Finally, we constrained the ancestral node for all rattlesnakes (i.e., *Crotalus* + *Sistrurus*), setting a normal distribution with an offset = 11.2 MYA, mean = 0, and SD = 3, following Reyes-Velasco et al. (2013). We ran two BEAST 2 runs for 5×10^9 generations each and evaluated effective sample sizes, stationarity, and run convergence. We discarded the first 10% of generations as burnin, combined estimates from both runs, and summarized parameter estimates on a maximum clade credibility tree using LogCombiner and TreeAnnotator v2.1.3 (Bouckaert et al. 2014).

We tested for evidence of simultaneous divergence of multiple lineages using the hierarchical approximate Bayesian computation (hABC) algorithm used in MTML-msBayes (Huang et al. 2011). Here, we specifically tested the hypothesis that divergence events of *C. scutulatus* and *C. atrox* populations across the Continental Divide were coincident. We prepared two inputs of lineage pairs: 1) Mojave-Sonoran + Chihuahuan *C. scutulatus* and western + eastern *C. atrox*. Using these inputs, we simulated 1,000,000 randomly drawn hyper-parameters and summary statistics and estimated the posterior density for numbers of possible divergence times, Ψ (i.e., 1-2), specifying a proportion of acceptable draws from the prior ('-t' option) to 0.005 (i.e., 500 draws from simulated priors).

Results

Mitochondrial gene phylogeny and genetic diversity

The 50% majority rule consensus phylogeny (Fig. 1A) of 46 unique *C. scutulatus* haplotypes revealed strong support (i.e., > 0.95 posterior) for *C. scutulatus* as monophyletic with respect to outgroup taxa, and for four distinct clades within *C. scutulatus*. The first major split within *C. scutulatus* was inferred as the divergence between two northern (Mojave-Sonoran and Chihuahuan) and two central/southern Mexico clades (Central Plateau and *salvini*). This split corresponds to the geographic break occurring in central Mexico, located where elevation increases in the Central Mexican Plateau by ~1,000 meters; samples from northern and southern Durango fall into northern and southern clades, respectively. We find evidence for two major subclades within each of these larger northern and southern clades (Fig. 1A). The major split within the northern clade segregates samples east and west of the Continental Divide (analogous to patterns observed in *C. atrox*; (Castoe et al. 2007; Schield et al. 2015)). The split within the

southern clade segregates samples from central and southern Mexican localities from samples in the known geographic range of *C. s. salvini* (Fig. 1B) (Campbell and Lamar 2004).

Marginal posterior probability densities for independent runs of IMa2 converged on consistent estimates of population genetic parameters (Table 1). The three analyses focusing on Mojave-Sonoran + Chihuahuan, Chihuahuan + Central Plateau, and Central Plateau + *salvini* mitochondrial clades, inferred similar estimates for parameters and scaled demographic units across replicates, and estimates reported below represent mean values from combined replicate analyses. Female population sizes were inferred to be greatest in the Mojave-Sonoran and Central Plateau clades (717,364 and 663,494 individuals, respectively), intermediate in the Chihuahuan clade (348,356 individuals), and two orders of magnitude lower in the southern Mexico *salvini* clade (4,489 individuals). We note, however, that the population size for the *salvini* population was estimated from a limited number of samples (n = 6) and should be regarded as tentative.

Mitochondrial genetic diversity was greatest in the Central Plateau and Chihuahuan populations (Table 2), with lower estimates in the Mojave-Sonoran population, and very low estimates in *salvini*. This pattern was consistent across different measures of genetic diversity and our estimates of population size, with the exception of haplotype diversity. Here, both the Mojave-Sonoran and Chihuahuan populations had greater haplotype diversity than the Central Plateau, yet this difference may also represent an artifact of ascertainment bias in haplotype sampling. We therefore consider the consistent patterns from within-population pairwise distances (π), and $\hat{\theta}$ to provide a more accurate depiction of the relative mitochondrial diversity across populations. Levels of expected versus observed mtDNA polymorphism resulted in negative *Fs* (consistent

with population expansion) for the Chihuahuan and Mojave-Sonoran populations (Table 2), yet coalescent simulations suggest that only the northern population had significant evidence of expansion at ($p < 0.02$; (Fu 1997)).

Nuclear estimates of phylogeny, population structure, and gene flow

The filtering scheme used for nuclear RAD loci resulted in 6,337 SNPs from 2,799 loci (Supplementary Table 1). Because we filtered to retain loci only if they were present in at least 30 of 44 individuals, the resulting data matrix had a relatively low proportion of missing data per individual (mean = 26.25%). Bayesian clustering analysis on nuclear SNPs (using a single, randomly sampled SNP per locus) in STRUCTURE provides consistent evidence for a best-fit model (ΔK) of three population clusters (Fig. 1C-D). Estimates of population assignments and admixture proportions from STRUCTURE runs were highly consistent across iterations. Under the $K = 3$ model, we find support for distinct Mojave-Sonoran and Chihuahuan genetic clusters, with Central Plateau and *salvini* samples assigned with high posterior probability to a single Central Plateau cluster. This model also provides implicit evidence of gene flow between populations based on shared posterior assignments of some individuals to multiple clusters – mixed-assignment individuals originate from localities adjacent to the neighboring population cluster, matching expectations of secondary contact between two distinct genetic lineages. Increasing values of K did not further segregate samples into unique genetic clusters, but instead identified small proportions of shared alleles among nearly all samples, potentially from shared ancestral variation. Three population clusters are also supported by DAPC analysis (which are independent of the model assumptions of STRUCTURE and tree-based analyses), where BIC identified a best-fit model of three distinct genetic clusters (Supplementary Fig. 1).

In contrast to the strong support for four distinct mtDNA clades, yet consistent with other analyses of nuclear loci, Bayesian nuclear phylogeny estimates support three distinct clades corresponding to major populations and regions throughout the range of *C. scutulatus* that coincide with inferences from population clustering analyses. While the three RADseq *C. s. salvini* samples were monophyletic, their subclade is nested within the clade comprising populations from the Central Plateau and the recognized range of *salvini* (i.e., south of central Durango, Mexico; Fig. 1C), which contrasts the unique and genetically distant mitochondrial clade of *salvini* endemic to the subspecies' proposed range. In our nuclear phylogenetic inferences, the Central Plateau + *salvini* clade had strong posterior support, and was inferred as sister lineage to the Mojave-Sonoran plus Chihuahuan clades. Further, the nuclear topology of Mojave-Sonoran and Chihuahuan clades was congruent with the mtDNA tree topology, splitting populations east and west of the Continental Divide, with additional structure in the Mojave-Sonoran population between subpopulations in California/Nevada and Arizona/New Mexico (Fig. 1C). Despite strong support from Bayesian tree inference on concatenated RAD loci for three distinct clades of *C. scutulatus* in which *salvini* is nested within the Central Plateau clade, Bayesian species delimitation analysis using bpp indicated 100% posterior probability for a four species model, comprising Mojave-Sonoran, Chihuahuan, Central Plateau, and *salvini* populations. Here, the best-supported unrooted topology matches the mtDNA topology (Supplementary Fig. 2).

We find moderate to high levels of nuclear genome-wide allelic differentiation between the three major nuclear clades, and patterns of differentiation are predicted by the mtDNA and nuclear phylogenies, with one exception. The highest F_{ST} is found between the Mojave-Sonoran

population and *salvini* (mean $F_{ST} = 0.197$), followed by the comparison between the Mojave-Sonoran population and the Central Plateau population ($F_{ST} = 0.158$), and a slightly lower F_{ST} (0.113) was observed between the Mojave-Sonoran and Chihuahuan populations. Interestingly, we observe lower differentiation between the Chihuahuan and Central-Plateau populations than between the Mojave-Sonoran and Chihuahuan populations ($F_{ST} = 0.098$), which we did not expect given the nuclear tree topology (Fig. 1D), but may reflect greater gene flow between populations adjacent to the phylogeographic break associated with the uplift of the Mexican Plateau. Alternatively, this inference may also be an artifact of the geographic distance between sampled localities on respective sides of the Continental Divide, resulting in a stronger signal of isolation-by-distance. Finally, we find allelic differentiation between *salvini* samples and the Central Plateau is very low ($F_{ST} = 0.026$).

Tests of Patterson's D statistics across both major phylogeographic breaks within the range of *C. scutulatus* (i.e., the Continental Divide and the uplift of the Mexican Plateau) revealed significant evidence of introgression, suggesting that these boundaries are permeable to gene flow between adjacent populations (Fig. 2). The 'Continental Divide' analysis resulted in a significantly positive value of D (0.2365, $p = 0.00024$), with 149 of 6,058 SNPs fitting the 'ABBA' pattern, and 92 'BABA' sites. Of the 5,077 SNPs used in the 'Mexican Plateau' analysis, there are 380 'ABBA sites' and 157 'BABA' sites, also resulting in a significantly positive D value (0.4153, $p < 0.00001$).

LGM modeling and historical biogeography

Ecological niche models derived from present-day environmental variables predicted suitable habitat that is largely overlapping with the known distribution of *C. scutulatus* (Fig. 3A); AUC

values for models under the four subsampling schemes ranged from 0.925 - 0.945, and we obtained an average AUC value of 0.934 for our combined model. The minimum training presence-absence threshold corresponded to a logistic probability of 0.05. We found that our subsampling scheme did not appreciably affect the results, and predictions were consistent across all 12 different reconstructions (Supplementary Fig. 3). The average LGM model (Fig. 3B) suggests that *C. scutulatus* has persisted within the Mojave and Sonoran Deserts and the southern Chihuahuan Desert into central Mexico during glacial periods. Within the Mojave-Sonoran region, populations would have been restricted such that their range in all directions was less expansive relative to the current northern distribution of the species. Likewise, the inferred LGM distribution within Mexico was also restricted in size and confined to the southern extent of the current predicted distribution, leaving a considerable gap in the LGM distribution between predicted Mojave-Sonoran and Chihuahuan + Central Plateau refugia (Fig. 3B). This inference is consistent with discontinuous suitable habitat during this and perhaps more ancient glacial periods. Fragmented LGM habitat in the northern extent of the ancestral *C. scutulatus* distribution could support one of two alternative hypotheses: 1) existing populations inhabiting this region were restricted in space and isolated from one another; partial or complete barriers to gene flow contributed to substantial divergence, or 2) the species occupied one of these refugia during the LGM and underwent subsequent range expansion into the region of the other inferred refugium when suitable habitat became connected. Estimates of population structure and divergence dating (see below) are consistent with the former (i.e., the split between these lineages predates the LGM), and evidence of geographic isolation between populations east and west of the Continental Divide during glacial periods provides historical and geographic context for their divergence. Additionally, we find evidence of discontinuity in predicted suitable habitat

in the extreme southern extent of the Mexican LGM distribution (Fig. 3B). The currently recognized distribution of *C. s. salvini* is restricted to this range, and potential isolation from adjacent southern populations may have facilitated spatial structuring of genetic diversity in this region.

Linear model tests of relationships between geography and nuclear genetic diversity revealed a pattern of spatial sorting of diversity consistent with greater diversity in the Chihuahuan and Central Plateau populations, and low levels of diversity in localities of the Mojave-Sonoran population consistent with range expansion (Fig. 3C-F). We found significant negative correlations between heterozygosity and both latitude ($N = 44$, $R^2 = -0.368$, $p = 0.0139$; Fig. 3C) and longitude ($N = 44$, $R^2 = -0.38$, $p = 0.0109$; Fig. 3D). We also found significant negative relationships between the number of private alleles and latitude ($N = 28$, $R^2 = -0.34$, $p = 0.0237$; Fig. 3E), as well as longitude ($N = 28$, $R^2 = -0.386$, $p = 0.0097$; Fig. 3F). These results imply an overall greater degree of genetic variation in the Chihuahuan and Central Plateau regions, and lower estimates with greater latitude and longitude are consistent with range expansion out of an ancestral range in central-southern Mexico. These estimates are also broadly consistent with patterns of mtDNA genetic diversity. With regard to the three samples from the currently recognized range of *C. s. salvini* (and corresponding with the mtDNA '*salvini*' clade), low levels of heterozygosity, but moderate numbers of private alleles, are the likely result of a population bottleneck and/or founder's effects during southern range expansion.

Divergence time estimates and evidence for lineage co-divergence

Parameter estimates across replicate BEAST runs were similar, with low variation in median divergence time estimates and sufficient ESS values (6,013 – 42,834). Median divergence time

estimates for focal clades range from 173 KYA for the southern *C. s. salvini* mitochondrial clade to 14.13 MYA for the split between *Crotalus* and *Agkistrodon* (Fig. 4A). Median estimates and 95% height posterior density distributions (HPD) for other divergence times among rattlesnake lineages correspond with estimates from Reyes-Velasco et al. (2013), with the exception of the common ancestor of *C. atrox* and the *C. scutulatus* group (median = 8.55 MYA), which is similar instead to estimates from Anderson and Greenbaum (2012). The median estimates and HPD for the splits between Mojave-Sonoran and Chihuahuan *C. scutulatus* (1.45 MYA; 0.78 – 2.5 MYA), and eastern *C. atrox* + western *C. atrox* (1.21 MYA; 0.76 – 1.75 MYA) overlap substantially, suggesting co-incident divergence events in these groups over the Continental Divide (Fig 4A). The median divergence time estimate for the split between the Central Plateau and *salvini* mitochondrial clades is more ancient, and while the HPD overlaps with the distributions from Mojave-Sonoran/Chihuahuan *C. scutulatus* and eastern/western *C. atrox*, this divergence event more closely corresponds with the relative age of the split between *C. ruber* and *C. atrox* (2.09 MYA; 1.14 – 3.49 MYA). Explicit inference of co-occurring divergence events within our dataset focused on divergence between populations adjacent to the Continental Divide (Mojave-Sonoran and Chihuahuan *C. scutulatus* and eastern and western *C. atrox*). The results from msBayes analysis indicated strong posterior support for synchronous divergence within these taxon pairs – the estimated dispersion index (Ω) heavily sampled 0 (mean = 0.0046, 95% HPD = 0.0 – 0.05; Fig. 4B). The simple rejection and categorical rejection methods for inferring the number of possible divergence times (Ψ) identified the greatest posterior probability support for $\Psi = 1$ (simple rejection *pp* = 0.821, rejection with multinomial logistic regression *pp* = 0.815).

Discussion

Evolutionary history and phylogeography of Crotalus scutulatus

Our analysis of population genetic variation across the range of *C. scutulatus* provides the first molecular evidence of substantial population genetic structure within this group of highly venomous snakes (Fig. 1). Our divergence time estimates suggest that *C. scutulatus* lineages diverged from a common ancestor with Western rattlesnakes during the Pliocene (roughly 3.4 MYA), and have subsequently diversified over an expansive geographic range. Our phylogenetic and population genetic inferences from mitochondrial and nuclear data generally support between three and four distinct lineages of *C. scutulatus* associated with relatively ancient divergences. Major lineages correspond with populations residing (from north to south) in the Mojave and Sonoran deserts in the United States, a large region of Chihuahuan Desert in Texas and Northern Mexico, higher elevation regions of the Mexican Plateau, and the extreme southern extent of the species range in Mexico, Puebla, and Veracruz, Mexico (the currently recognized range of *C. s. salvini*; Fig. 1).

We find consistent evidence from both mtDNA and nuclear SNPs for two major phylogeographic breaks within *C. scutulatus*: the Continental Divide and the uplift of the Mexican Plateau. The 'Cochise Filter Barrier' region surrounding the Continental Divide is a biogeographic zone of restricted gene flow between numerous taxa (Morafka 1977b), including multiple snake species (Myers et al. 2016). Thus, *C. scutulatus* divergence at this region further implicates this region as an important vicariant barrier among the North American desert taxa. Restriction of populations to suitable habitat on either side of this region during the Pleistocene likely promoted divergence between lineages adjacent to the Continental Divide (Fig. 3) until

more recent secondary contact after range expansion from both the Mojave-Sonoran and Chihuahuan regions (Table 2).

The second major phylogeographic break in *C. scutulatus* identified by our results appears to be associated generally with the uplift of the Mexican Plateau, which represents a transition between lowland Chihuahuan desert habitat and higher elevation semi-arid habitat, and it is also a well-studied biogeographic barrier (Morafka 1977a; Marshall and Liebherr 2000; Bryson et al. 2011). Our results suggest that Pleistocene climatic cycles did not contribute to *C. scutulatus* diversification across this barrier, as we infer suitable habitat during glacial periods to have broad overlap with both the Chihuahuan and Central Plateau regions (Fig 3B). Thus, we expect limits to gene flow across the uplift of the Mexican Plateau to have been driven by ecological constraints associated with the transition from Chihuahuan desert habitat to the central Mexican Matorral, rather than changes in suitable habitat driven by climatic fluctuations. This prediction is also supported by the divergence time estimate between the ancestral lineages of extant *C. scutulatus* predating Pleistocene glaciation.

We observed a diversification pattern in *C. scutulatus* across the Continental Divide that was similar to our previous findings for *C. atrox* (Castoe et al. 2007; Schield et al. 2015; Schield et al. 2017), and found consistent evidence for the synchronous divergence of eastern and western *C. atrox* lineages and Mojave-Sonoran and Chihuahuan *C. scutulatus* (Fig. 4A-B) at this boundary. This finding suggests that common historical processes during the Pleistocene drove simultaneous divergence of these taxa across this boundary. In a recent study, Myers et al. (2016) investigated diversification across the Continental Divide in twelve species and found broad evidence for asynchronous diversification sufficient to reject the hypothesis of a single vicariant

event among all taxa studied. With respect to *C. scutulatus* and *C. atrox,* however, they observed overlapping posterior divergence time distributions, consistent with our present findings. Combined, these results support synchronous Pleistocene diversification for multiple rattlesnake lineages, suggesting that taxa with similar ecologies, life histories, and climatic affinities (like *C. atrox* and *C. scutulatus*) are more likely to co-diverge across barriers due to shared biotic and abiotic factors. In addition to the Continental Divide, there are several rattlesnake species other than *C. scutulatus* with distributions that span the uplift of the Mexican Plateau barrier. A previous broad biogeographical study of both Mojave-Sonoran and Chihuahuan Deserts based on various taxa showed congruent patterns and timing of diversification across lineages in this area (Flores-Villela and Martinez-Salazar 2009). These findings suggest that further comparative studies testing for coincident divergence at the Mexican Plateau barrier would be valuable for assessing if this biogeographic feature also resulted in detectable co-divergence in multiple rattlesnake species (or even multiple snake species).

While mtDNA and nuclear loci both highlight previously unappreciated lineage diversity within *C. scutulatus* and agree in many inferences of lineage diversification within this taxon, they disagree with regard to the relative placement and distinctiveness of populations currently allocated to *C. s. salvini.* Both our mitochondrial phylogeny and nuclear-based species delimitation inferences strongly support four distinct clades of *C. scutulatus* (Fig. 1A, Fig. 4A). However, our concatenated RADseq phylogeny and population clustering analyses support only three distinct clades in which *C. s. salvini* is nested within the Central Plateau lineage. Furthermore, genome-wide F_{ST} is very low between *salvini* and the Central Plateau populations, consistent with either pervasive gene flow, very recent divergence, or both. Thus, while we find

substantial evidence for three distinct lineages of *C. scutulatus*, evidence for a fourth distinct lineage (representing *C. s. salvini*) is conflicting across datasets and analyses.

Below we discuss two likely factors that may explain the discrepancies between analyses that favor either three or four lineages, both of which suggest that the three-lineage model is most likely – accordingly, we discuss species hypotheses under a three-lineage model hereafter. First, mito-nuclear discordance is a well-documented phenomenon with several potential explanations (Sloan et al. 2016), including biases inherent to the mode of mitochondrial inheritance and incomplete lineage sorting (Degnan and Rosenberg 2009). Because of the recent divergence of these clades, we expect gene tree conflicts due to incomplete lineage sorting to be pervasive across our datasets. Another common driver of mito-nuclear discordance is sex-biased gene flow (Toews and Brelsford 2012). *Crotalus scutulatus* is one of several rattlesnake species that demonstrate high levels of male-biased dispersal (males disperse greater than five times farther than females when reproductively active; (Cardwell 2008). Thus, we expect that mito-nuclear discordance has been driven by biased dispersal of males compared to that of females, leading to the combination of a completely endemic mtDNA haplotype and including pervasive nuclear gene flow. Second, although bpp analyses suggest four distinct clades matching the mtDNA topology, we find evidence of gene flow between populations, violating the core assumptions of the biological species concept implemented in the model (Yang and Rannala 2010). In this sense, we suspect that nuclear genetic structure in the region of endemic *C. s. salvini* mtDNA at most represents incipient speciation in the early stages of divergence with gene flow. Alternatively, this signal could be driven by isolation during glacial periods (Fig. 3) followed by secondary

contact where male-biased gene flow has deteriorated nuclear genetic structure with the persistence of mitochondrial structure.

Species hypotheses and evidence for gene flow between divergent lineages

Two subspecies of *C. scutulatus* are currently recognized: the Mojave Rattlesnake (*C. s. scutulatus*) and the Huamantlan Rattlesnake (*C. s. salvini*; (Campbell and Lamar 2004). Under the current classification, the Mojave Rattlesnake occupies the vast majority of the northern portion of the species' distribution, while the Huamantlan Rattlesnake is restricted to a smaller, high elevation (i.e., > 1,800 meters; (Campbell and Lamar 2004) region at the southern end of the species' distribution. Our results suggest that *C. s. salvini* and the Central Plateau population are synonymous (*C. s. salvini*, hereafter), and that the distribution of *C. s. salvini* in fact extends north to the central Mexican Plateau. Furthermore, our results provide evidence for cryptic, taxonomically unrecognized lineage diversity within *C. scutulatus*, as phylogenetic and population genetic analyses of both mtDNA and nuclear loci consistently support a distinct third 'Chihuahuan' lineage that geographically ranges between the Continental Divide and the Mexican Plateau.

The divergence time estimate between *C. s. salvini* and the common ancestor of Mojave-Sonoran and Chihuahuan *C. scutulatus* (3.4 MYA; Fig. 4A) is more ancient than several recognized rattlesnake species pairs (e.g., *C. ruber* and *C. atrox*, 2.99 MYA (Reyes-Velasco et al. 2013); *C. viridis* and *C. oreganus*, 3.1 MYA (Anderson and Greenbaum 2012). Additionally, the divergence between the Mojave-Sonoran and Chihuahuan *C. scutulatus* (1.45 MYA; Fig. 4A) is also more ancient than some recognized rattlesnake species (e.g., *C. durissus* and *C. simus*, ~1

MYA (Blair and Sanchez-Ramirez 2016). Based on this precedent, these three *C. scutulatus* lineages represent excellent 'species hypotheses' to be tested.

It is becoming clear from molecular studies of rattlesnakes that considerable divergence (i.e., millions of years) does not necessarily prevent gene flow and introgression in situations of secondary contact (Murphy and Ben Crabtree 1988; Castoe et al. 2007; Meik et al. 2015; Schield et al. 2015). For example, divergent populations of the Western Diamondback Rattlesnake (*C. atrox*) east and west of the Continental Divide introgress over a broad geographic region in the northern Chihuahuan Desert (Schield et al. 2017). Divergent lineages of *C. scutulatus* also exhibit this pattern, with evidence of individuals with admixed assignment from Bayesian population clustering analysis in STRUCTURE (Fig. 1B) and a significantly positive value of Patterson's D (Fig. 2A). These findings also highlight the apparent present-day permeability of major historic barriers to gene flow. For example, although the Continental Divide has been repeatedly demonstrated as an important historical driver of diversification in many arid-adapted taxa, migrants from divergent rattlesnake lineages are able to introgress across this barrier. Evidence for gene flow across the uplift of the Mexican Plateau between Chihuahuan and Central Plateau populations is also notable, given that these lineages diverged from an even more ancient ancestor than did the lineages that meet at the Continental Divide. Patterns of gene flow in *C. scutulatus* therefore add to a growing body of evidence that distinct lineages (even species) may experience substantial levels of admixture or hybridization in secondary contact (Murphy and Ben Crabtree 1988; Zancolli et al. 2016; Schield et al. 2017), underscoring rattlesnakes as an important model system for understanding incomplete reproductive isolation and the evolution (or lack thereof) of postzygotic isolation mechanisms in secondary contact.

Conclusion

In this study, we investigated range-wide patterns of diversity, population structure, and gene flow in the Mojave Rattlesnake. Through analyses of both mtDNA and nuclear genomic data we find consistent evidence for previously undocumented lineage diversity within this group. These lineages represent intriguing species hypotheses to be further tested using complementary data to the genomic data presented here (e.g., morphological and venom variation data). Our findings also underscore the importance of using multiple types of genetic markers to explore range-wide diversity in such wide-ranging lineages, which in our case led to the identification of multiple apparently distinct lineages and of important physiographic features promoting divergence. Beyond biogeographic and taxonomic implications, previously undocumented diversity within *Crotalus scutulatus* has immediate medical relevance due to the toxicity and diversity of its venom composition. Evidence of population genetic structure (and gene flow) presented here may thus provide a useful framework for detailed comparative and evolutionary analyses of venom variation in general, and specifically for understanding the apparently complex gain and loss of neurotoxic function in this broadly distributed species complex (Glenn and Straight 1978; Glenn et al. 1983; Glenn and Straight 1989).

Acknowledgments

We thank Jonathan Campbell, Carl Franklin, Luis Felipe Vázquez-Vega, Edmundo Pérez-Ramos, Jesús Sigala, Jens Vindum (California Academy of Sciences), and Corey Roelke for providing tissue samples; Elda Sanchez, Mark Hockmuller, and Juan Salinas for assistance with tissue samples from the Texas A&M Kingsville Natural Toxins Research Center. Scientific collecting permits were issued by the California Department of Fish and Wildlife (SC-12985),

the New Mexico Department of Game and Fish (3563, 3576), the State of Arizona Game and Fish Department (SP628489, SP673390, SP673626, SP715023), Texas Parks and Wildlife (SPR-0390-029) and Secretaria de Medio Ambiente y Recursos Naturales of the Estados Unidos Mexicanos (SGPA/DGVS/03562/15). Support for this work was provided by faculty startup funds from the University of Texas at Arlington to TAC, NSF Grant DEB-1655571 to TAC, SPM, and JMM, UC Mexus CN-11-548 to CLS, NSF DDIG Grant DEB-1501886 awarded to DRS and TAC, NSF DDIG Grant DEB-1501747 to DCC and TAC, a Phi Sigma Beta Phi Chapter research grant to DRS, and a Theodore Roosevelt Memorial Fund research grant, Prairie Biotic Research Inc. Grant, Sigma Xi Grants-in-aid-of-research, SnakeDays Research Grant, and Southwestern Association of Naturalists Howard McCarley research grant to JLS.

Figures

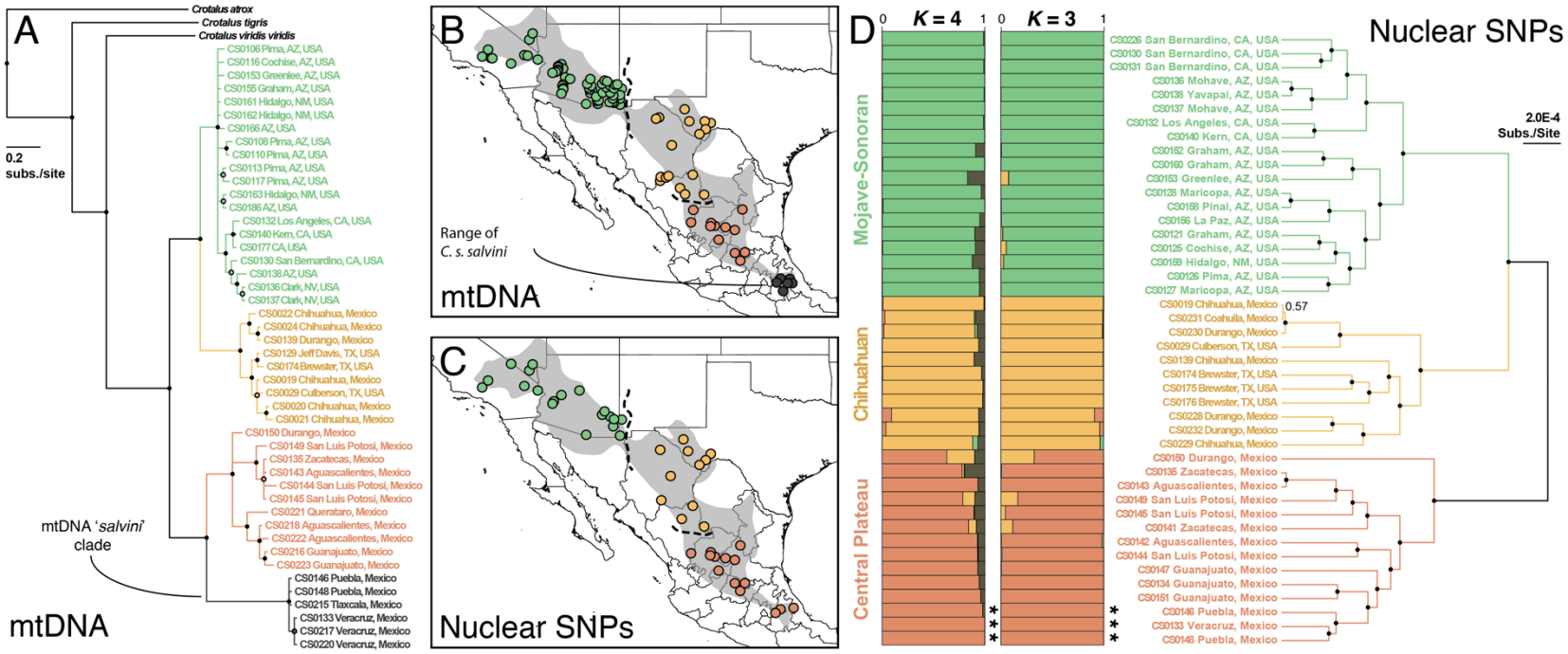


Figure 1. Mitochondrial and nuclear evidence for phylogenetic and population structure across the range of *C. scutulatus*. **A.** Bayesian mitochondrial gene (ND4) tree depicting four distinct clades, with sample identifiers and general locality descriptions provided. **B.** Map of sampling localities used in mtDNA analyses; dots are colored according to mitochondrial clade (the current range of *C. s. salvini* is indicated). Dashed lines represent the Continental Divide (northern) and uplift of the Mexican Plateau (southern) biogeographic barriers. **C.** Sampling localities used nuclear SNP analyses; dots are colored according to major nuclear clade. The range of *C. scutulatus* shaded in gray on each map was redrawn from spatial data available from the IUCN. **D.** Bayesian population clustering models from STRUCTURE and phylogenetic tree (based on concatenated RAD loci). Horizontal bars on $K = 3$ and 4 plots depict the posterior probability of assignment to one or more ancestral population clusters, colored to correspond with mtDNA/nuclear trees. Samples from the currently recognized range of *C. s. salvini* are marked with asterisks. On each tree, posterior support > 0.95 for each node is summarized by a black dot and support > 0.8 is represented by an open circle. Posterior support for the shallow node between samples CS0019 and CS0231 is specifically labeled.

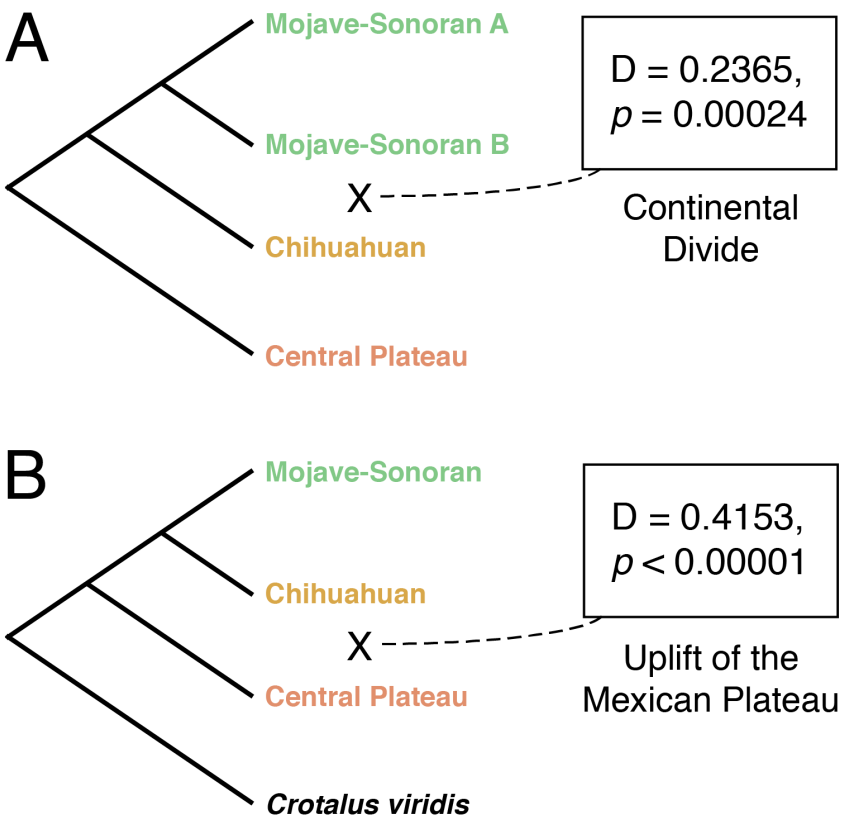


Figure 2. Schematic design and results of ABBA-BABA tests using Patterson's D statistics for introgression across biogeographic barriers. **A.** Tree topology used to test for introgression across the Continental Divide and D-statistic information. **B.** Tree topology used to test for introgression across the uplift of the Mexican Plateau, including *C. viridis* as outgroup, and information about the resulting D-statistic. Colors used to label the *C. scutulatus* tree tips in these schematics correspond with the nuclear tree colors in Fig. 1.

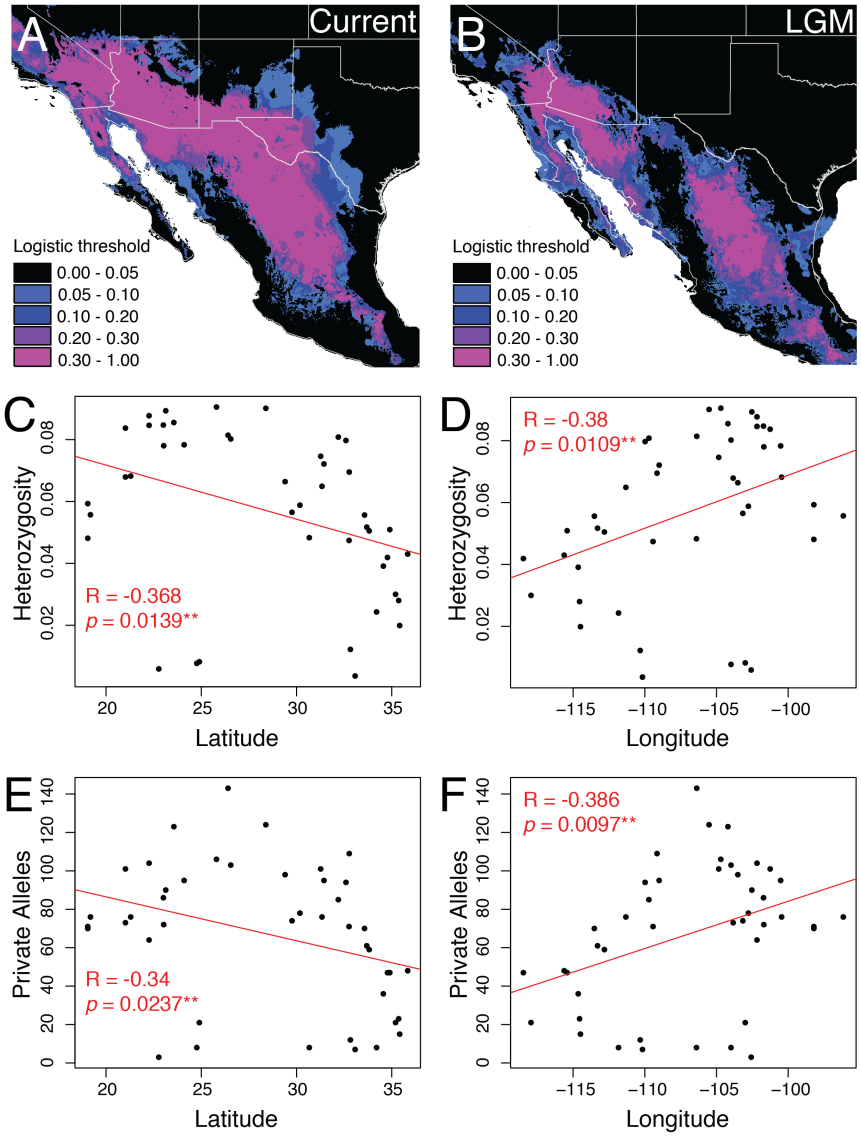


Figure 3. Inferred current and historical suitable habitat based on ecological niche modeling of *C. scutulatus* localities and relationships between nuclear genetic diversity and geography. **A-B.** Current and Last Glacial Maximum projections, respectively, of suitable habitat for *C. scutulatus* estimated using niche modeling. Warmer colors depict regions with a high logistic threshold probability of suitable habitat, while dark colors represent areas with low probability of *C. scutulatus* presence. **C-D.** Scatterplots of estimated heterozygosity versus latitude and longitude, respectively. **E-F.** Scatterplots of numbers of private alleles versus latitude and longitude, respectively. Black dots represent individuals and trendlines, R-values, and *p* values are shown in red for each linear model.

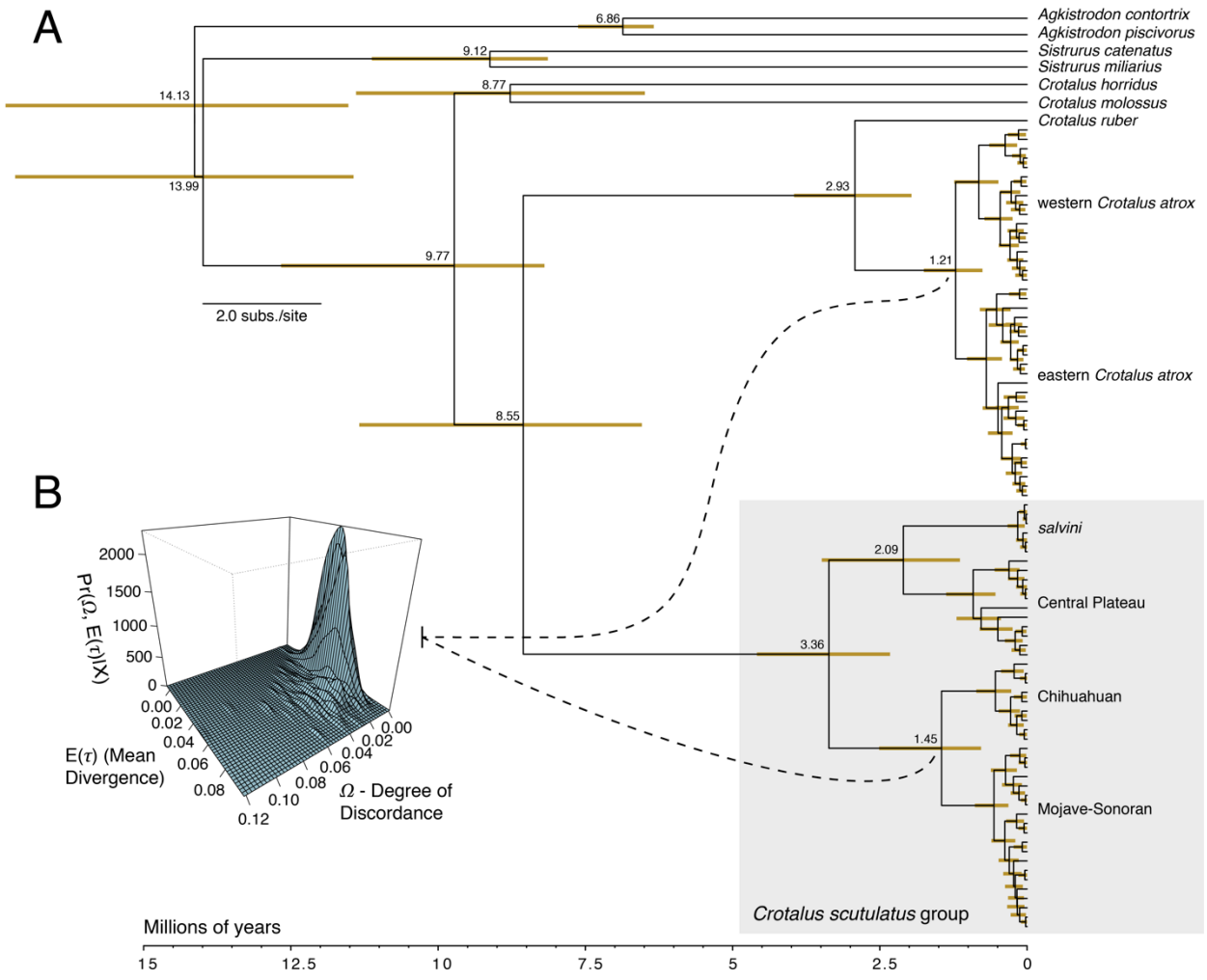


Figure 4. Divergence time estimates and evidence for synchronous divergence of rattlesnake populations across the Continental Divide at the 'Cochise Filter Barrier'. **A.** BEAST time tree based on analysis of mtDNA, showing *C. scutulatus*, *C. atrox*, and outgroup relationships and divergence time estimates, with the *C. scutulatus* group shaded in grey. Median estimates of node ages for major divergence events are labeled on the tree, and yellow bars at each node represent the 95% HPD. A scale of time in millions of years is shown at the bottom. **B.** Results of tests of co-divergence in msBayes, depicting the combined distributions of the degree of discordance (Ω), mean relative divergence time (t), and probability distribution of mean divergence time and discordance given alignments of Mojave-Sonoran and Chihuahuan *C. scutulatus* and eastern and western *C. atrox*.

Table 1. Population genetic parameter estimates and scaled demographic values for three parallel analyses of the isolation-migration model. 'MLE' is equivalent to the highest peak of the posterior density distribution.

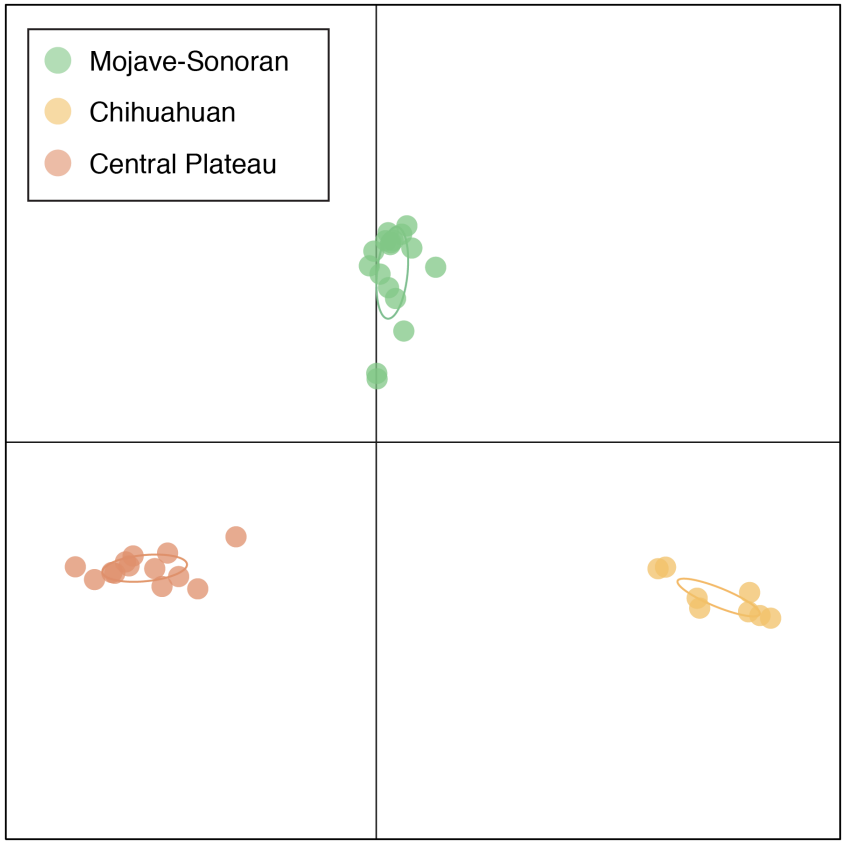
	(1) Mohave-Sonoran + (2) Chihuahuan		(1) Chihuahuan + (2) Central Plateau		(1) Central Plateau + (2) *salvini*	
	MLE	95% HPD	MLE	95% HPD	MLE	95% HPD
θ_1	39.95	24.55 / 63.35	19.95	7.85 / 51.05	39.35	18.95 / 89.05
θ_2	18.85	7.45 / 48.35	34.55	16.45 / 74.55	0.25*	0 / 13.45
θ_A	17.75	0 / 94.95	48.95	5.45 / 99.95	58.35	5.95 / 99.95
t	9.75	5.25 / 16.25	20.75	13.75 / 29.75	17.57	0 / 49.98
N_{e1}	717,364	440,833 / 1,137,547	358,233	140,958 / 916,681	706,590	340,276 / 1,599,030
N_{e2}	338,480	133,776 / 868,199	620,398	295,385 / 1,338,660	4,489*	0 / 241,515
N_{eA}	-	0 / 1,704,974	-	32,440 / 594,940	-	106,841 / 1,794,757
T_{MRCA}	1,050,420	565,610 / 1,750,700	2,235,510	1,481,362 / 3,205,128	1,892,911	0 / 5,384,615

*Parameters likely influenced by small sample size in *salvini*. We do not provide the estimates for ancestral population sizes, as the posterior density curves for these parameters were flat, and high point estimates are therefore not meaningful. Abbreviations: θ_1; unscaled population 1 size parameter, θ_2; unscaled population 2 size parameter, θ_A; unscaled inferred ancestral population size parameter, *t*; unscaled time since divergence, N_{e1}; effective female population 1 size, N_{e2}; effective female population 2 size, N_{eA}; effective inferred ancestral female population size, T_{MRCA}; time to most recent common ancestor (years).

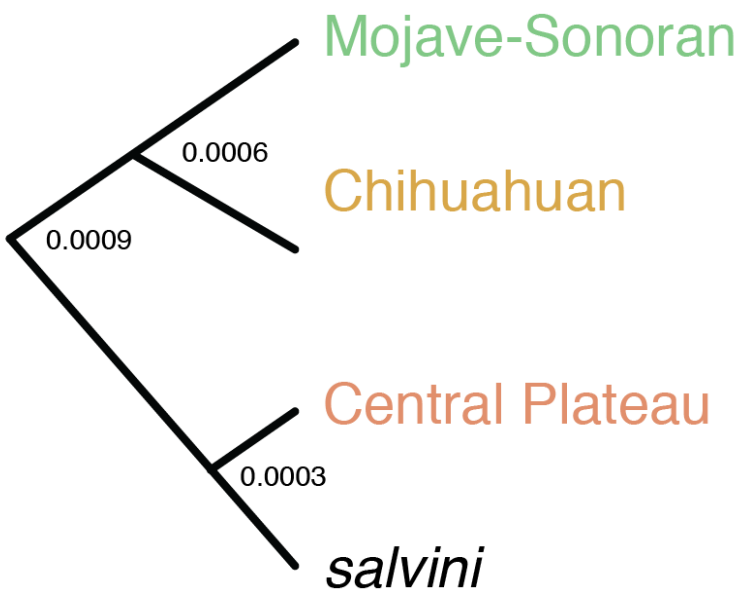
Table 2. Measures of nucleotide diversity and tests for evidence of population expansion within *C. scutulatus* mtDNA clades.

Population	Pairwise Distance	Haplotype Diversity	Pi	Theta	Fu's Fs	p-value
Mohave-Sonoran	0.0687	0.963	0.005	3.221	-7.967	< 0.00001*
Chihuahuan	0.085	0.987	0.0083	5.889	-2.278	0.089
Central Plateau	0.1085	0.905	0.013	9.818	2.225	0.26
salvini	0.0212	0.6	0.0008	0.6	0.795	0.8

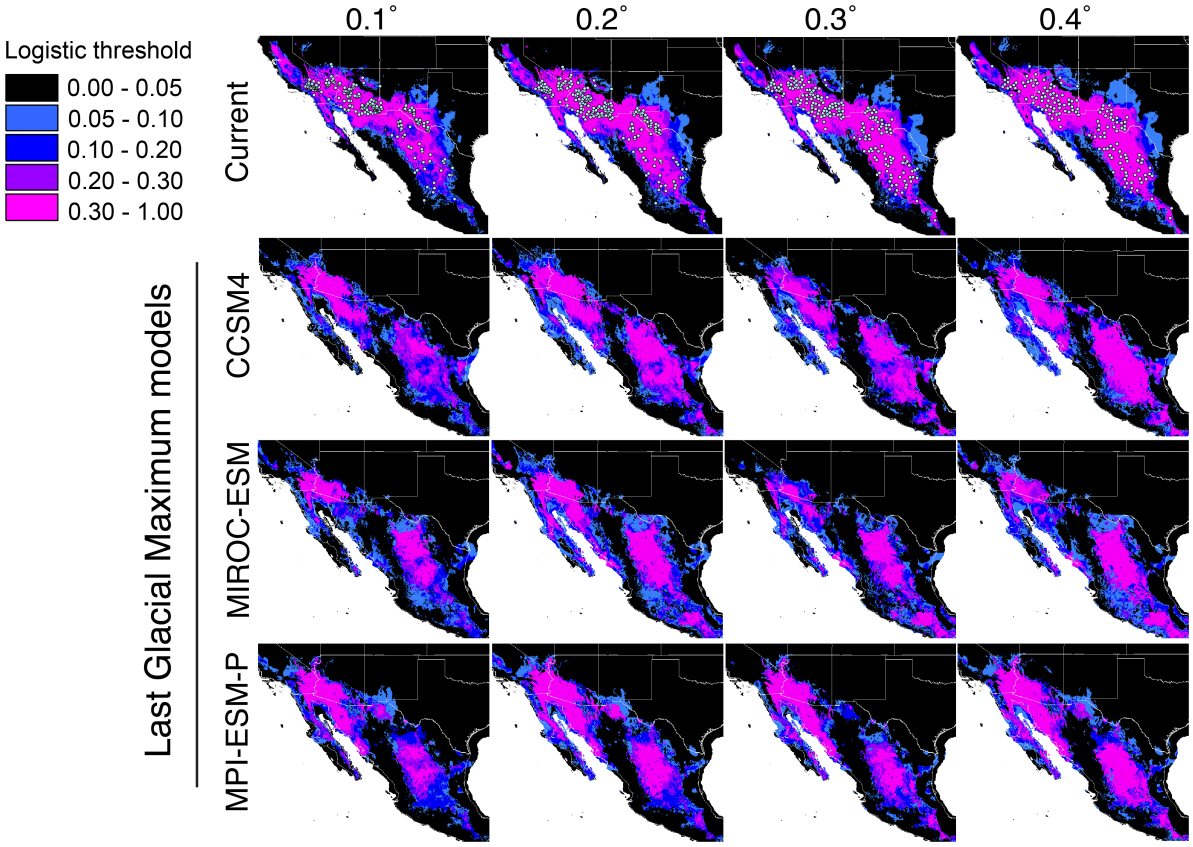
SUPPLEMENTARY FIGURES



Supplementary Figure 1. Results of Discriminant Analysis of Principal Components (DAPC), displaying BIC-supported three-cluster model. Colors of clustered genetic groups correspond to colors used in Fig. 1 and Fig. 2 in the main text.



Supplementary Figure 2. Results of species delimitation in bpp based on nuclear RADseq loci, supporting a four-species model (see text). Values at the nodes are relative branch lengths, and the tips are colored according to the lineage designations in Fig. 1 in the main text.



Supplementary Figure 3. Current and historical *C. scutulatus* habitat predictions from ENM analysis. Shaded regions represent regions of suitable habitat with various logistic threshold probabilities. The top row of panels represent inferred current suitable habitat. Green dots represent *C. scutulatus* localties. The remaining rows are projections of suitable habitat onto conditions of the Last Glacial Maximum climatic models that were averaged and are presented in Fig. 3. The four columns of projections are based on the four occurrence record subsampling schemes meant to reduce potential bias from greater sampling in the United States, setting a minimum distance between samples of 0.1°, 0.2°, 0.3°, and 0.4°, respectively.

SUPPLEMENTARY TABLE

Supplementary Table 1. Specimen data for samples used in this study. Where noted, samples were used previously in Castoe et al. (2007) and Schield et al. (2015).

Species	Project ID	Museum ID	Voucher/other ID	Country	State	County	Reference
Crotalus scutulatus	CS0156		SPM307	USA	Arizona	La Paz	this study
Crotalus scutulatus	CS0159		SPM318	USA	New Mexico	Hidalgo	this study
Crotalus scutulatus	CS0130	CAS 228094		USA	California	San Bernardino	this study
Crotalus scutulatus	CS0158	CAS 170520		USA	Arizona	Pinal	this study
Crotalus scutulatus	CS0160	CAS 170452		USA	Arizona	Graham	this study
Crotalus scutulatus	CS0144		NMB056	Mexico	San Luis Potosi		this study
Crotalus scutulatus	CS0147		Rob Bryson 6	Mexico	Guanajuato		this study
Crotalus scutulatus	CS0142		NMB037	Mexico	Aguascalientes		this study
Crotalus scutulatus	CS0126		Csx005a	USA	Arizona	Pima	this study
Crotalus scutulatus	CS0125		Csx004	USA	Arizona	Cochise	this study
Crotalus scutulatus	CS0129	CAS 70519		USA	Arizona	Jeff Davis	this study
Crotalus scutulatus	CS0133		CLS881	Mexico	Veracruz		this study
Crotalus scutulatus	CS0121		Csx001	USA	Arizona	Graham	this study
Crotalus scutulatus	CS0139	MZFC 17996	KWS294	Mexico	Chihuahua		this study
Crotalus scutulatus	CS0137		CLS909	USA	Arizona	Mohave	this study
Crotalus scutulatus	CS0146		OFV1124	Mexico	Puebla		this study
Crotalus scutulatus	CS0136		CLS908	USA	Arizona	Mohave	this study
Crotalus scutulatus	CS0140	MVZ 137600		USA	California	Kern	this study
Crotalus scutulatus	CS0141	MZFC 22927	ANMO 1369	Mexico	Zacatecas		this study
Crotalus scutulatus	CS0134		CLS891	Mexico	Guanajuato		this study
Crotalus scutulatus	CS0132		CLS811	USA	Californa	Los Angeles	this study
Crotalus scutulatus	CS0029	NTRC 345		USA	Texas	Culberson	this study
Crotalus scutulatus	CS0127		DRS0019	USA	Arizona	Maricopa	this study

Crotalus scutulatus	CS0019	UTA R 58937	JAC29013	Mexico	Chihuahua		this study
Crotalus scutulatus	CS0174	ASNHC15001	CLP1980	USA	Texas	Brewster	this study
Crotalus scutulatus	CS0175	ASNHC15002	CLP2021	USA	Texas	Brewster	this study
Crotalus scutulatus	CS0128		DRS0020	USA	Arizona	Maricopa	this study
Crotalus scutulatus	CS0224	UTA R 58939	JAC29080	Mexico	Chihuahua		this study
Crotalus scutulatus	CS0176	ASNHC15003	CLP2027	USA	Texas	Brewster	this study
Crotalus scutulatus	CS0145		NMB058	Mexico	San Luis Potosi		this study
Crotalus scutulatus	CS0151		NMB074	Mexico	Guanajuato		this study
Crotalus scutulatus	CS0149		DAR108	Mexico	San Luis Potosi		this study
Crotalus scutulatus	CS0228		TP30852	Mexico	Durango		this study
Crotalus scutulatus	CS0150	MZFC 25003	JMM 242	Mexico	Durango		this study
Crotalus scutulatus	CS0229	MVZ 226656		Mexico	Chihuahua		this study
Crotalus scutulatus	CS0131	MVZ 265259	CLS800	USA	California	San Bernardino	this study
Crotalus scutulatus	CS0143		NMB038	Mexico	Aguascalientes		this study
Crotalus scutulatus	CS0148		CLS868	Mexico	Puebla		this study
Crotalus scutulatus	CS0138		CLS911-RB4	USA	Arizona	Yavapai	this study
Crotalus scutulatus	CS0152	MVZ 265428	PFS0263	USA	Arizona	Graham	this study
Crotalus scutulatus	CS0153	MVZ 265429	PFS0265	USA	Arizona	Greenlee	this study
Crotalus scutulatus	CS0226	MVZ 229838		USA	California	San Bernardino	this study
Crotalus scutulatus	CS0135		CLS892	Mexico	Zacatecas		this study
Crotalus scutulatus	CS0227		CLS915	USA	California	San Bernardino	this study
Crotalus scutulatus	CS0230	—	MBCS005	Mexico	Durango		this study
Crotalus scutulatus	CS0231	—	MBCS006	Mexico	Coahuila		this study
Crotalus scutulatus	CS0232	—	MBCS007	Mexico	Durango		this study
Crotalus scutulatus	CS0234	—	XPR049	Mexico	Mexico		this study
Crotalus scutulatus	CS0020		JAC29014	Mexico	Chihuahua		this study
Crotalus scutulatus	CS0021	UTA R 58938	JAC29015	Mexico	Chihuahua		this study
Crotalus scutulatus	CS0022		JAC29076	Mexico	Chihuahua		this study
Crotalus scutulatus	CS0024	UTA R 58940	JAC29089	Mexico	Chihuahua		this study

Crotalus scutulatus	CS0027	NTRC 926		USA	Arizona	Pima	this study
Crotalus scutulatus	CS0028	NTRC 928		USA	Arizona	Pima	this study
Crotalus scutulatus	CS0030	NTRC 924		USA	Arizona	Cochise	this study
Crotalus scutulatus	CS0077		CS005b	USA	Arizona	Pima	this study
Crotalus scutulatus	CS0081		CS009	USA	Arizona	Pinal	this study
Crotalus scutulatus	CS0084		CS012	USA	Arizona	Pima	this study
Crotalus scutulatus	CS0086		CS014	USA	Arizona	Pima	this study
Crotalus scutulatus	CS0090		CS018	USA	Arizona	Pima	this study
Crotalus scutulatus	CS0091		CS019	USA	Arizona	Pima	this study
Crotalus scutulatus	CS0092		CS020	USA	Arizona	Pima	this study
Crotalus scutulatus	CS0093		CS021	USA	Arizona	Pima	this study
Crotalus scutulatus	CS0105		CSPR001	USA	Arizona	Pima	this study
Crotalus scutulatus	CS0106		CSPR003	USA	Arizona	Pima	this study
Crotalus scutulatus	CS0107		CSPR004	USA	Arizona	Pima	this study
Crotalus scutulatus	CS0108		CSPR005	USA	Arizona	Pima	this study
Crotalus scutulatus	CS0110		CSPR007	USA	Arizona	Pima	this study
Crotalus scutulatus	CS0111		CSPR008	USA	Arizona	Pima	this study
Crotalus scutulatus	CS0112		CSPR012	USA	Arizona	Pima	this study
Crotalus scutulatus	CS0113		CSPR013	USA	Arizona	Pima	this study
Crotalus scutulatus	CS0114		CSPR014	USA	Arizona	Pima	this study
Crotalus scutulatus	CS0116		CSPR016	USA	Arizona	Cochise	this study
Crotalus scutulatus	CS0117		CSPR017	USA	Arizona	Pima	this study
Crotalus scutulatus	CS0118		CSPR018	USA	Arizona	Pima	this study
Crotalus scutulatus	CS0119		CSPR019	USA	Arizona	Cochise	this study
Crotalus scutulatus	CS0120		CSPR020	USA	Arizona	Pima	this study
Crotalus scutulatus	CS0122		Csx002	USA	Arizona	Cochise	this study
Crotalus scutulatus	CS0123		Csx0021	USA	Arizona	Cochise	this study
Crotalus scutulatus	CS0124		Csx003	USA	Arizona	Cochise	this study
Crotalus scutulatus	CS0155	CAS 170506		USA	Arizona	Pima	this study

Crotalus scutulatus	CS0157	CAS 170451		USA	Arizona	Pinal	this study
Crotalus scutulatus	CS0161		PFS0320	USA	New Mexico	Hidalgo	this study
Crotalus scutulatus	CS0162		PFS0322	USA	New Mexico	Hidalgo	this study
Crotalus scutulatus	CS0163		PFS0335	USA	New Mexico	Hidalgo	this study
Crotalus scutulatus	CS0164	ASNHC14996	CLP1929	USA	New Mexico	Hidalgo	this study
Crotalus scutulatus	CS0165	ASNHC14997	CLP1930	USA	New Mexico	Hidalgo	this study
Crotalus scutulatus	CS0166	ASU36035	CLP1936	USA	Arizona	Graham	this study
Crotalus scutulatus	CS0167	ASU36075	CLP1953	USA	Arizona	Maricopa	this study
Crotalus scutulatus	CS0169	ASU36061	CLP1959	USA	Arizona	Yavapai	this study
Crotalus scutulatus	CS0170	ASU36062	CLP1961	USA	Arizona	Yavapai	this study
Crotalus scutulatus	CS0171	ASU36077	CLP1963	USA	Arizona	Maricopa	this study
Crotalus scutulatus	CS0172	ASU36091	CLP1971	USA	Arizona	Pima	this study
Crotalus scutulatus	CS0173	ASU36092	CLP1972	USA	Arizona	Pima	this study
Crotalus scutulatus	CS0177	CAS259909	CLP2075	USA	California	Los Angeles	this study
Crotalus scutulatus	CS0178	CAS259910	CLP2076	USA	California	Los Angeles	this study
Crotalus scutulatus	CS0179	CAS259911	CLP2077	USA	California	San Bernardino	this study
Crotalus scutulatus	CS0180	ASU36078	CLP2096	USA	Arizona	Maricopa	this study
Crotalus scutulatus	CS0181	CAS259916	CLP2108	USA	California	San Bernardino	this study
Crotalus scutulatus	CS0182	ASU36063	CLP2111	USA	Arizona	Santa Cruz	this study
Crotalus scutulatus	CS0183	ASU36036	CLP2112	USA	Arizona	Graham	this study
Crotalus scutulatus	CS0184	ASU36064	CLP2113	USA	Arizona	Santa Cruz	this study
Crotalus scutulatus	CS0185	ASU36065	CLP2114	USA	Arizona	Santa Cruz	this study
Crotalus scutulatus	CS0186	ASU36066	CLP2115	USA	Arizona	Santa Cruz	this study
Crotalus scutulatus	CS0187	ASU36067	CLP2116	USA	Arizona	Santa Cruz	this study
Crotalus scutulatus	CS0188	ASU36103	CLP2136	USA	Arizona	Pinal	this study
Crotalus scutulatus	CS0189	ASU36104	CLP2142	USA	Arizona	Pinal	this study
Crotalus scutulatus	CS0190	ASU36039	CLP2152	USA	Arizona	Graham	this study
Crotalus scutulatus	CS0191	ASU36037	CLP2153	USA	Arizona	Graham	this study
Crotalus scutulatus	CS0192	ASU36038	CLP2154	USA	Arizona	Graham	this study

Crotalus scutulatus	CS0193	ASU36040	CLP2166	USA	Arizona	Cochise	this study
Crotalus scutulatus	CS0194	ASU36041	CLP2167	USA	Arizona	Cochise	this study
Crotalus scutulatus	CS0195	ASU36042	CLP2168	USA	Arizona	Cochise	this study
Crotalus scutulatus	CS0196	ASNHC14998	CLP2172	USA	New Mexico	Hidalgo	this study
Crotalus scutulatus	CS0197	ASNHC14999	CLP2173	USA	New Mexico	Hidalgo	this study
Crotalus scutulatus	CS0198	ASU36043	CLP2174	USA	Arizona	Cochise	this study
Crotalus scutulatus	CS0199	ASU36044	CLP2175	USA	Arizona	Cochise	this study
Crotalus scutulatus	CS0200	ASNHC15000	CLP2182	USA	New Mexico	Hidalgo	this study
Crotalus scutulatus	CS0201	ASNHC15004	CLP2191	USA	Texas	Brewster	this study
Crotalus scutulatus	CS0202	—	CLPT162	USA	Arizona	Yuma	this study
Crotalus scutulatus	CS0203	—	CLPT167	USA	Arizona	Maricopa	this study
Crotalus scutulatus	CS0204	—	CLPT172	USA	Arizona	Pinal	this study
Crotalus scutulatus	CS0205	—	CLPT173	USA	Arizona	Pinal	this study
Crotalus scutulatus	CS0207	—	CLPT177	USA	Arizona	Maricopa	this study
Crotalus scutulatus	CS0208	—	CLPT180	USA	Arizona	Yavapai	this study
Crotalus scutulatus	CS0209	—	CLPT181	USA	Arizona	Yavapai	this study
Crotalus scutulatus	CS0210	—	CLPT195	USA	New Mexico	Hidalgo	this study
Crotalus scutulatus	CS0211	—	CLPT200	USA	New Mexico	Hidalgo	this study
Crotalus scutulatus	CS0213		CLS 903	Mexico			this study
Crotalus scutulatus	CS0214		CLS 893	Mexico	Zacatecas		this study
Crotalus scutulatus	CS0215		CLS 902	Mexico	Tlaxcala		this study
Crotalus scutulatus	CS0216		NMB 075	Mexico	Guanajuato		this study
Crotalus scutulatus	CS0217		CLS 876	Mexico	Veracruz		this study
Crotalus scutulatus	CS0218		CLS 896	Mexico	Aguascalientes		this study
Crotalus scutulatus	CS0219	CAS 170505		USA	Arizona		this study
Crotalus scutulatus	CS0220		CLS 875	Mexico	Veracruz		this study
Crotalus scutulatus	CS0221		CLS 905	Mexico	Querataro		this study
Crotalus scutulatus	CS0222		CLS 895	Mexico	Aguascalientes		this study
Crotalus scutulatus	CS0223		CLS 899	Mexico	Guanajuato		this study

Crotalus scutulatus	CS0233	—	MBCS008	Mexico	Coahuila		this study
Crotalus scutulatus	CS0235	—	XPR050	Mexico	Mexico		this study
Crotalus atrox	CA0006		CLS474	USA	Arizona	Cochise	Castoe et al. 2007
Crotalus atrox	CA0012		RNF2595	USA	California	Imperial	Castoe et al. 2007
Crotalus atrox	CA0013		RNF2596	USA	California	Imperial	Castoe et al. 2007
Crotalus atrox	CA0028		RWV2001-22	USA	Texas	Jeff Davis	Castoe et al. 2007
Crotalus atrox	CA0031		ENS10537	Mexico	San Luis Potosi		Castoe et al. 2007
Crotalus atrox	CA0032		ENS10538	Mexico	Veracruz		Castoe et al. 2007
Crotalus atrox	CA0033		ENS10539	Mexico	Zacatecas		Castoe et al. 2007
Crotalus atrox	CA0034		ENS10540	Mexico	Zacatecas		Schield et al. 2015
Crotalus atrox	CA0035		ENS10536	Mexico			Castoe et al. 2007
Crotalus atrox	CA0042		ENTA21	USA	Arizona	Pinal	Castoe et al. 2007
Crotalus atrox	CA0043		ENTA49	USA	Arizona	Pinal	Castoe et al. 2007
Crotalus atrox	CA0047		ENT12	USA	New Mexico	Hidalgo	Castoe et al. 2007
Crotalus atrox	CA0048		ENT7	USA	Arizona	Maricopa	Castoe et al. 2007
Crotalus atrox	CA0049	LACM150957		USA	Arizona	Pima	Castoe et al. 2007
Crotalus atrox	CA0051		CLP60	USA	Texas	El Paso	Castoe et al. 2007
Crotalus atrox	CA0052		CLP64	USA	Texas	Jeff Davis	Castoe et al. 2007
Crotalus atrox	CA0058		TJL588	USA	Texas	Goliad	Castoe et al. 2007
Crotalus atrox	CA0069		TJL348	USA	Texas	Val Verde	Castoe et al. 2007
Crotalus atrox	CA0072		JJ	USA	Texas	Dallas	Castoe et al. 2007
Crotalus atrox	CA0075		TWR 1249	Mexico	Sonora		Castoe et al. 2007
Crotalus atrox	CA0082	ROM18144		USA	California	Riverside	Castoe et al. 2007
Crotalus atrox	CA0083	ROM18244		Mexico	Baja		Castoe et al. 2007
Crotalus atrox	CA0085		JAC 26689	USA			Schield et al. 2015
Crotalus atrox	CA0086	UTA R 58967	JAC 29236	USA			Castoe et al. 2007
Crotalus atrox	CA0087	UTA R 58968	JAC 29282				Castoe et al. 2007
Crotalus atrox	CA0088	UTA R 57694	JAC 29568	USA			Schield et al. 2015
Crotalus atrox	CA0091	UTA R 58966	JAC 29854	USA			Castoe et al. 2007

Crotalus atrox	CA0098		RLG380	USA	Texas	LaSalle	Schield et al. 2015
Crotalus atrox	CA0100		RLG404	USA	Texas	Zavala	Schield et al. 2015
Crotalus atrox	CA0106		CLS378	USA	New Mexico	Hidalgo	Schield et al. 2015
Crotalus atrox	CA0132		CLS244	USA	New Mexico	Hidalgo	Schield et al. 2015
Crotalus atrox	CA0149		CLS286	USA	New Mexico	Hidalgo	Castoe et al. 2007
Crotalus atrox	CA0151		CLS290	USA	Arizona	Cochise	Castoe et al. 2007
Crotalus atrox	CA0160		CLS300	USA	New Mexico	Hidalgo	Castoe et al. 2007
Crotalus atrox	CA0164		CLS304	USA	Arizona	Cochise	Castoe et al. 2007
Crotalus atrox	CA0183		CLS354	USA	New Mexico	Hidalgo	Schield et al. 2015
Crotalus atrox	CA0191		CLS368	USA	Arizona	Cochise	Schield et al. 2015
Crotalus atrox	CA0193		DRS0002	USA	Texas	Shackelford	Schield et al. 2015
Crotalus atrox	CA0194		DRS0003	USA	Texas	PaloPinto	Schield et al. 2015
Crotalus atrox	CA0197		JWS656	USA	Texas	Jeff Davis	Schield et al. 2015
Crotalus viridis	CV0017		SPM-316	USA	New Mexico	Dona Ana	this study
Crotalus viridis	CV0007		SPM-W08152	USA	Colorado	Weld	this study
Crotalus viridis	CV0317	NTRC 636		USA	Nebraska	Dawes	this study
Crotalus viridis	CV0320	NTRC 237		USA	Oklahoma	Woods	this study
Crotalus ruber							
Crotalus molossus			CLP66				
Crotalus horridus			CSL8758				
Sistrurus miliarius		ROM18232					
Sistrurus catenatus							
Agkistrodon piscivorus							
Agkistrodon contortrix							

Chapter 5

A chromosome-level prairie rattlesnake genome provides new insight into reptile genome biology and gene regulation in the venom gland

Drew R. Schield[1], Daren C. Card[1], Nicole R. Hales[1], Blair W. Perry[1], Giulia M. Pasquesi[1], Heath Blackmon[2], Richard H. Adams[1], Andrew B. Corbin[1], Balan Ramesh[1], Jeffrey P. Demuth[1], Marc Tollis[3], Jesse M. Meik[4], Stephen P. Mackessy[5], and Todd A. Castoe[1, §]

[1]Department of Biology & Amphibian and Reptile Diversity Research Center, University of Texas at Arlington, Arlington, TX, USA

[2]Department of Biology, Texas A&M University, College Station, TX, USA

[3]School of Life Sciences, Arizona State University, Tempe, AZ, USA

[c]Department of Biological Sciences, Tarleton State University, 1333 W. Washington Street, Stephenville, TX, 76402 USA

[5]School of Biological Sciences, University of Northern Colorado, Greeley, CO, USA

Abstract

Here we present the genome sequence of the prairie rattlesnake (*Crotalus viridis viridis*), the first chromosome-level vertebrate genome generated using only next-generation sequencing, and use this genome to study key biological features of reptiles, and venomous snakes specifically. We identify the full length rattlesnake Z chromosome, including the recombining pseudoautosomal region, and demonstrate remarkable similarities in Z chromosome evolution and structure between snake and avian species. We also find evidence for incomplete dosage compensation, and identify multiple mechanisms that appear to contribute to the incomplete dosage on the snake Z chromosome. This genome also provides some of the first clear insight into the origins, structure, and function of reptile microchromosomes, which we find have markedly different structure and function compared to macrochromosomes. Rattlesnake microchromosomes harbor elevated gene density, show substantial variation in regional GC content, and (based on analysis of the 3D chromatin structure) appear to interact with other chromosomes at a much higher frequency than do macrochromosomes. This genome assembly also allowed, for the first time, identification of the chromosomal locations of all rattlesnake venom gene families. We find that microchromsomes are particularly enriched for venom genes, which we show have evolved through multiple tandem duplication events of multiple gene families. By overlaying 3D chromatin structure information and gene expression data we identify specific transcription factors that direct expression on venom genes, and demonstrate how chromatin structure guides precise expression of multiple venom gene families. Together, analyses of the prairie rattlesnake genome reveal multiple key features of reptile genome biology, and provide insight into the origins, structure, and regulation of a complex and dynamic phenotype - snake venom.

Introduction

Snakes have become important model systems for understanding the evolution of specialized and extreme phenotypes, such as limblessness (Cohn and Tickle 1999), metabolic adaptation (Castoe et al. 2008), and a spectrum of adaptations linked to prey acquisition. Indeed, snakes possess several of the most striking examples of adaptations for feeding known in vertebrates, including a highly kinetic skull allowing snakes to feed on large prey (Gans 1961), extreme physiological and metabolic fluctuations in response to feeding (Secor and Diamond 1998), and the evolution of toxic venoms for prey capture together with highly-specialized structures for storing and delivering venom (i.e., venom glands and fangs). Snakes and other squamate reptiles have also played increasingly prominent roles in studies of genomic repeat element evolution (Castoe et al. 2011; Castoe et al. 2013; Pasquesi et al. In Review), GC isochore structure (Fujita et al. 2011; Castoe et al. 2013), and the evolution of sex chromosomes (Matsubara et al. 2006; Vicoso et al. 2013). In particular, snakes are a valuable system for understanding the evolutionary trajectories of sex chromosome evolution because snakes have evolved both ZW and XY sex chromosomes independently several times (Gamble et al. 2017), and snake species exhibit a spectrum of sex chromosome differentiation (Matsubara et al. 2006), ranging from karyologically indistinguishable homomorphic (e.g., in boas), to highly heteromorphic differentiated sex chromosomes in vipers.

Limited genomic resources and fragmented genome assemblies have been a barrier to fully leveraging snakes as model systems for studying the genomic basis of extreme adaptations and the evolution genome structure (Bradnam et al. 2013; Castoe et al. 2013; Vonk et al. 2013; Yin et

al. 2016). To address this, we constructed a high-quality genome of the prairie rattlesnake (*Crotalus viridis viridis*) using a combination of high-throughput sequencing and Hi-C scaffolding (Lieberman-Aiden et al. 2009). The prairie rattlesnake is a pitviper native to North America, which possesses potent and complex venom. This species, and viperid snakes in general, have highly-differentiated sex chromosomes and remarkable genome-wide variation in transposable element abundance and diversity. Here we use the rattlesnake genome, which is the first chromosome-level genome assembly for a reptile, to address multiple hypotheses regarding snake and vertebrate genome evolution, and to provide new insight into the regulatory mechanisms underlying venom production in the rattlesnake venom gland.

Results and Discussion

A chromosome-level rattlesnake genome

We sequenced and assembled the genome of a male Prairie Rattlesnake (*Crotalus viridis viridis*) at 1,658-fold physical coverage using multiple high-throughput sequencing approaches combined with the Dovetail Genomics HiRise sequencing and assembly method (Supplementary Tables 1 and 2), which combines long-range Chicago data (Rice et al. 2017) with 3D chromatin contact information from Hi-C. This approach resulted in the most contiguous reptile genome to date (*CroVir3.0*), with a scaffold N50 of 179.9 Mbp, represented by 3 scaffolds. We estimate the total genome size to be between 1.25 and 1.34 Gbp based on k-mer frequency distributions and the assembled genome size, respectively. The genome annotation contains 17,352 predicted protein-coding genes, and an annotated repeat element content of 39.49% (Supplementary Table 3).

The rattlesnake is the first vertebrate genome to achieve chromosome-level assembly using only next-generation sequencing technology, and the first ever snake chromosome-level assembly, including all chromosomes in the rattlesnake karyotype (2n = 36). Chromosome identities of large scaffolds were further confirmed using chromosome-specific gene markers (Matsubara et al. 2006), which mapped uniquely to large scaffolds corresponding to macrochromosomes (Chromosomes 1 through 7; Supplementary Table 4). Microchromosomes were originally over-assembled into a single large scaffold, which was manually split based on multiple lines of evidence (see below and Supplementary Methods). The corrected assembly resulted in microchromosome scaffolds with lengths matching the size predictions of the rattlesnake karyotype (Baker et al. 1972). Finally, we identified the rattlesnake Z chromosome using multiple lines of evidence, which we discuss further below. The chromosomal sequences include assembled telomeric and centromeric regions, with centromeres containing an abundant 164 bp monomer with 42% GC content (Supplementary Fig. 1).

This chromosome-level assembly provides new insight into the genome-wide distribution of key features and homology across amniote genomes. In the rattlesnake, we find the microchromosomes to contain the highest and most variable GC content. We also find rattlesnake microchromosomes have particularly high gene density (100 Kb windows; p-values < 0.00001) and reduced repeat element content compared to macrochromosomes ($p < 0.00001$; Fig. 1a), similar to patterns observed in the Chicken (Supplementary Fig. 2). Rattlesnake chromosomes show high degrees of synteny with those from *Anolis*, except for an apparent fusion/separation of *Anolis* chromosome 3 into snake chromosomes 4 and 5 (Fig. 1b). Microchromosomes also appear largely homologous across squamates (although this is limited

by resolution of microchromosome linkage groups in *Anolis*). We were also able to further validate previous inferences that *Anolis* chromosome 6 is homologous to the sex chromosomes of rattlesnakes. Despite conservation of squamate microchromosome homology, patterns of chicken-squamate homology suggest that there have been major shifts between macro- and microchromosome locations for large syntenic regions of the genome. The chicken has a large number of microchromosomes, and we find that about half of these are syntenic with squamate microchromosomes (Fig. 1b), but that other microchromosomes are syntenic with blocks of squamate macrochromosomes (i.e., squamate chromosome 2 is an amalgam of chicken microchromosomes and the Z chromosome). We also observe large regions of synteny between chicken and squamate macrochromosomes. For example, squamate chromosome 1 shares large syntenic tracks with chicken chromosomes 3, 5, and 7 (Fig. 1b). Surprisingly, the largest chicken macrochromosomes (1 and 2) show synteny patterns that are scattered across multiple squamate macrochromosomes, including the rattlesnake Z chromosome, indicating multiple exchanges of genomic regions between macro- and microchromosomes early in amniote evolution.

Squamate reptiles have become particularly important for studying the evolution of genomic GC content and isochore structure, due to the loss of GC isochores in *Anolis* yet the apparent re-emurgence of isochore structure in snakes (Fujita et al. 2011; Castoe et al. 2013). To visualize genomic GC variation, we compared orthologous aligned genomic regions across 12 squamates, which demonstrates that there have been two major transitions in genomic GC content, including a reduction in GC content from lizards to snakes, and a secondary further reduction in GC content within the colubroid lineage of snakes that includes the rattlesnake and cobra (Fig. 1c). This suggests that higher genome-wide GC content was likely the ancestral squamate condition,

and that snakes have evolved increased GC variation through an increase in genomic AT content (i.e., AT isochores), rather than a buildup of GC-rich islands, as was suggested by the finding of AT-biased substitutions from lizard to python and cobra genomes (Castoe et al. 2013). Interestingly, the negative relationship between genomic GC content (Fig. 1c) and GC isochore structure across squamate evolution (Fig. 1d; Supplementary Table 8) further suggests that GC-biased gene conversion cannot explain GC variation in snakes – a finding that has broad ramifications for understanding the mechanisms underlying shifts in genomic nucleotide content and variation, and the spatial structure of this variation. In addition to genomic GC content variation, squamate reptiles are notable because they appear to have remarkably variable, active, and rapidly-evolving genomic repeat element content across lineages, which is surprising given the relatively small and conserved genome size of squamates (Castoe et al. 2011; Pasquesi et al. In Review), and our analyses here confirm these trends (Fig. 1e). We find the genomes of colubroid snakes are dominated largely by several DNA elements (e.g., hAT and Tc1) and by non-LTR retrotransposons, and CR1-L3 LINEs in particular. The rattlesnake genome, specifically, has the highest abundance of CR1-L3s among sampled colubroid genomes (Fig. 1e), and low divergence of the majority of rattlesnake CR1-L3s suggests that these elements are quite active in the genome (Supplementary Fig. 3).

Sex chromosome evolution mechanisms of dosage compensation

The contiguity of the rattlesnake genome facilitates new perspectives into the structure of the pseudoautosomal region (PAR) and evolutionary strata of snake sex chromosomes, as well as new information on patterns of dosage compensation in snakes that provide new parallels across amniotes for understanding sex chromosome evolution. Recent studies have shown that snake

sex chromosomes have evolved multiple times, apparently from different autosomal chromosomes (Gamble et al. 2017), and have suggested that colubroid Z/W chromosomes are homologous with *Anolis* chromosome 6 (Srikulnath et al. 2009; Vicoso et al. 2013). We identified a single 114 Mb scaffold as the rattlesnake Z chromosome, which was confirmed by its broad synteny with *Anolis* chromosome 6, the presence of multiple known Z-linked markers ((Matsubara et al. 2006); Supplementary Table 4), and with coverage of mapped genomic reads that match expectations of hemizosity based on additional genomic data we collected from female individuals (Matsubara et al. 2006); Fig. 2a; Supplementary Fig. 4). We further identified the Z/W recombining PAR as the distal 7.2 Mb region of the Z chromosome that shows equal male-female genomic read depth (Fig. 2a) – this rattlesnake PAR is GC-rich relative to the genomic background and the non-PAR Z-chromosome regions (42.9%; Supplementary Fig. 5), similar to the pattern observed in the PAR of the Collard Flycatcher (Smeds et al. 2014). This suggests that common processes (e.g., GC-biased gene conversion) may drive increased GC content in the recombining regions of the independently evolved snake and avian sex chromosomes. The rattlesnake PAR also exhibits distinctive patterns of repeat element content compared to the Z, with lower levels of divergence among particular repeat elements in the PAR (e.g., CR1 and Bov-B LINEs), suggesting more recent element activity and insertion (Supplementary Fig. 6). We also find higher gene density in the rattlesnake PAR than elsewhere in the Z chromosome (Fisher's Exact Test: $p = 4.46 \times 10^{-7}$; Supplementary Fig. 7).

The existence of evolutionary strata on snake sex chromosomes has been suggested (Vicoso et al. 2013; Yin et al. 2016), but prior analyses have lacked the important context of a contiguous Z chromosome assembly. In addition to the PAR (i.e., Stratum 3), we identified a secondary

evolutionary stratum situated between the PAR and the remaining Z chromosome. This region (Stratum 2) shows near-autosomal levels of female-male ratios of mapped genomic reads (Fig. 2a,b). We hypothesize based on its location between the PAR and the oldest recombination-suppressed region (Stratum 1), combined with observed intermediate female:male read depth, that Stratum 2 represents a recombination-suppressed region that has retained substantial homology between Z and W chromosomes (Fig. 2b). Consistent with this hypothesis, a comparison of within-individual nucleotide diversity between females and males revealed elevated diversity in the female across Stratum 2, likely explained by the mapping of reads to divergent Z and W-linked paralogs in Stratum 2 in females (Fig. 2a,b). This suggests that a number of W-linked gene copies have been retained over the course of W chromosome degeneration and divergence from the Z chromosome, as has been hypothesized for birds (Bellott et al. 2017). The oldest evolutionary stratum (Stratum 1) is characterized by half female coverage relative to that of males, and roughly zero female nucleotide diversity, consistent with female hemizogosity across Stratum 1 (Fig. 2b). Based on our Hi-C data, we also find that the evolutionary strata on the Z chromosome broadly coincide with the boundaries of inferred topologically-associated domains (TADs; Fig. 2a), which provides the first precise demonstration that chromatin organization co-evolves with recombination suppression and sex chromosome differentiation.

Dosage compensation in organisms with differentiated sex chromosomes is of broad interest, especially due to the surprising diversity of mechanisms by which dosage is accomplished (Graves 2016). Colubroid snakes have been shown to exhibit partial dosage compensation (Vicoso et al. 2013; Yin et al. 2016), yet no mechanisms for compensation have been proposed.

The absence of complete dosage compensation is also supported by our data, which demonstrate that the overall ratio of female:male gene expression is significantly lower on the Z-chromosome compared to that of autosomes ($p < 10^{-16}$; Fig. 2c; Supplementary Fig. 8). Intriguingly, we identified patterns of partial or incomplete dosage compensation that varied widely across regions of the Z-chromosome, ranging from a total lack of compensation to equal expression in females and males (Fig. 2a and c), further raising the question of what mechanisms drive such variation.

To address mechanisms that might underlie partial compensation, we analyzed gene expression data from males and females for two different tissues (liver and kidney) in a stratum-specific fashion. In Stratum 3, we find that gene expression ratios between sexes largely match those on autosomes for both tissues (Fig. 2c; liver $p = 0.366$, kidney $p = 0.453$). This further confirms the identification of this region as the PAR, where compensation is achieved by the Z and W being homologous and effectively autosomal. In Stratum 2, in addition to intermediate female:male genomic read coverage, we find evidence for intermediate dosage (Fig. 2c), consistent with partial dosage compensation in females due to what we hypothesize represents effective diploidy through retained W-linked Z-chromosome homologs. Indeed, 24.5% of genes with female:male expression ratios greater than the 5th quantile of autosomal female:male ratio are within this region; these genes combined with genes in the PAR constitute 46.9% of dosed genes on the Z. Therefore, our results suggest that a substantial proportion of 'dosage' is driven by effective diploidy in females for genes in Strata 2-3. Finally, Stratum 1 showed the most variation in dosage, ranging from nearly complete to absent. We tested for evidence of a female-biased transcriptional regulatory mechanism (estrogen response elements; EREs) that could explain

regional or gene-specific compensation, and find that this mechanism may only account (at best) for a small number (8.5%) of dosed Stratum 1 genes, which we estimated were linked (i.e., within 100 Kb) to a predicted ERE (Fig. 2a). These findings suggest that additional unidentified dosage compensation mechanisms likely exist in snakes, which may include post-transcriptional mechanisms, as have been implicated in partial chicken dosage compensation (Uebbing et al. 2015).

Hi-C reveals unique microchromosome biology

Our analyses of the first available 3D chromatin contacts for a non-mammalian vertebrate (Fig. 3a) provide new perspectives on high-order genome organization and contact structure in reptiles and unique features of microchromosome biology. Patterns of intra- and interchromosomal chromatin contacts across rattlesnake macrochromosomes are broadly consistent with patterns observed in mammals, such that when interchromosomal contact frequencies are normalized by chromosome length, they show a consistent negative linear relationship across species (Fig. 3b). However, rattlesnake microchromosomes show a much steeper negative slope, deviating significantly from expectations based on macrochromosome contact frequencies. These data indicate an unexpected higher degree of contact between microchromosomes and other chromosomes (Fig. 3a), and a surprisngly high degree of interchromosomal contact among microchromosomes (Fig. 3c). In fact, the initial misassembly of microchromosomes into a single scaffold was likely driven by unexpected high frequencies of contact among microchromosomes, which significantly exceed assumptions of genome assembly based on mammalian macrochromosomes ($t = 13.38$, $p < 2.2 \times 10\text{-}16$, Fig. 3d) – this assembly error was later corrected using complementary information from chromatin contact frequencies, reptile microchromosome

synteny, and patterns of GC and repeat content (Supplementary Methods). Importantly, this first demonstration of Hi-C assembly of microchromosomes indicates that similar steps may need to be taken in future Hi-C sequencing and assembly projects for organisms with microchromosomes, and highlights the uniqueness of microchromosome interactions within the nucleus of at least snakes, if not other amniotes.

Microchromosomes are present in most birds and reptiles, but tend to be poorly represented and characterized in existing assembled genomes. Further, much of what we understand about microchromosome biology comes from studies of birds, and limited comparisons with other species (e.g., *Anolis* (Alfoldi et al. 2011) and *Pogona* (Georges et al. 2015) lizards) suggest that genomic features of microchromosomes may differ among species, despite the existence of considerable reptile microchromosome synteny (Fig. 1b). A comparison of compositional features between micro- and macrochromosomes of other species suggests that the rattlesnake exhibits patterns remarkably similar to chicken and zebra finch (i.e., significantly higher GC and gene content and lower repeat content on microchromosomes than on macrochromosomes), with the exception of higher repeat content in zebra finch microchromosomes (Supplementary Fig. 2). Lizards are more variable, with lower gene density in microchromosomes than rattlesnake and the birds (microchromosome gene density in *Pogona* is lower than in macrochromosomes), however, consistencies among species possessing microchromosomes suggest that ancestral amniote microchromosomes likely exhibited patterns similar to those observed in both the rattlesnake and chicken (Fig. 1a, Fig. 3a-c, Supplementary Fig. 2).

Insight into the origins, evolution, and regulation of snake venom and its production

Snake venoms and venom systems are intriguing examples for studying the evolution of biological novelty and represent topics of intense study and medical relevance (Mackessy 2010; Arnold 2016). The rattlesnake genome provides the first clear insight into the genomic location, organization, and broader genomic context for snake venom gene family evolution (Fig 4a). Our localization of rattlesnake venom genes to chromosomes revealed that venom gene families are enriched for being located on microchromosomes ($p = 0.0017$). Moreover, microchromosome-linked families include three of the most abundant, well-characterized, and medically-relevant components of prairie rattlesnake venom (Fig. 4a; snake venom metalloproteinases, SVMPs; snake venom serine proteinases, SVSPs; and type IIA phospholipases A2, PLA2s) – each of these families is located on a different microchromosome. The only remaining major component of prairie rattlesnake venom, myotoxin (crotamine), is located on Chromosome 1 (Fig. 4a). The intriguing location and abundance of venom genes on microchromosomes suggests intimate associations between microchromosome biology and venom evolution. To identify the origins and mechanisms underlying the evolution of these venom families we conducted phylogenetic estimates of each of the microchromosome-linked families listed above (including non-venom members) and inferred that each venom family represents a distinct set of tandemly-duplicated genes derived from a single duplication that gave rise to a monophyletic cluster of venom paralogs (Fig. 4b). While this mechanism has been proposed previously (Ikeda et al. 2010; Vonk et al. 2013), the contiguituy of our genome assembly provides the first definitive proof of this representing a repeated mechanism underlying the origin of snake venom gene clusters.

Using gene expression data from multiple venom gland samples and a diversity of other tissues, we find that genes in these venom clusters can be further readily demarcated by their distinctive venom gland-specific expression patterns (Fig. 5), which also highlights marked expression differences between the venom cluster versus flanking non-venom genes of PLA2, SVMP, and SVSP gene families. Such discrete expression patterns of adjacent venom and non-venom genes raises the intriguing question of how venom genes are uniquely regulated and targeted for expression in venom glands.

To understand mechanisms underlying venom-gland-targeted expression of venom genes we combined Hi-C, gene expression, and genome information, and took advantage of the fact that snake venom glands are paired (one on the left, one on the right side). First, to investigate the chromatin architecture of venom production, we extracted venom from one venom gland of the genome animal two days prior to the other gland, then dissected the venom glands one day after the second gland was extracted – this accomplished staggering the process of venom expression in these two glands, providing a 1 and 3 day post-extraction design. Hi-C sequencing of the 1-day post-extraction venom gland then enabled us to capture the chromatin contacts underlying venom production, which we further investigated by comparing gene expression between the two glands. Based on our Hi-C data, we find that the precise genomic regions containing venom clusters show a highly specific chromatin structuresituated within discrete high-frequency contact regions representing distinct topologically-associated domains (TADs; (Dixon et al. 2016) of open chromatin (Fig. 5). Genes adjacent to, and outside of these venom-specific TADs exhibit significantly lower expression in the venom gland, indicating a remarkably strong

insulating regulatory effect of TAD boundaries surrounding venom cluster regions (Fig. 5b), which also serves to block the spread of positive regulators to non-venom regions.

To identify transcription factors that may be responsible for directing venom-gland specific expression of venom genes, we compared gene expression levels of all annotated transcription factors in the rattlesnake genome between venom glands and other tissues. Here, we specifically tested for significant evidence that transcription factors exhibit an expression profile similar to the observed profiles for the venom gene clusters (Fig. 5a). This analysis identified a set of candidate transcription factors of interest that were significantly more highly expressed in the venom gland, including six with specific DNA binding function: *FOXC2, SREBF2*, and four members of the *CTF/NFI* family of DNA-binding transcription factors (*NFIA*, two isoforms of *NFIB,* and *NFIX*; Supplementary Table 5). To narrow this candidate set of putative venom-driving transcrion factors, we tested for evidence that predicted binding site sequences for these transcription factors were over-represented specifically in venom genes. We find that the upstream regions of genes in each of the three main venom clusters are significantly enriched for predicted *NFI* transcription factor binding sites (p-values < 0.05), but not *FOXC2* or *SREBF2* (p-values > 0.05). The combination of venom gland expression specificity and binding site enrichment analyses thus imply a central role of the *NFI* family of transcription factors in regulating expression of snake venom. Additionally, *NFI* has low binding affinity for nucleosomal DNA (Chikhirzhina et al. 2008), and the inferred open chromatin state within venom clusters should further enable efficient binding of highly-expressed *NFI* isoforms to their predicted binding sites, indicating the important complementary roles of both regional open

chromatin state and venom-gland-specific transcription factors in the regulation of snake venom production.

While not directly involved in venom gene regulation, we also find evidence that other transcription factors that exhibit significant upregulation in the venom gland play important roles in in venom production, such as those involved in the unfolded protein response of the endoplasmic reticulum (e.g., *ATF6* and *CREB3L2*) and in glandular epithelial development and maintenance (e.g., *ELF5* and *GRHL1*). While not immediately obvious, the increased activity of each of these categories of transcription factor in the venom gland makes sense considering the distinct features and requirements of venom production. For example, the rapid and immediate production of venom proteins following the release of venom (Luna et al. 2009) is expected to place incredibly high demands (and stress) on the endoplasmic reticulum as proteins are packaged and secreted into the venom gland lumen, and increased expression of factors involved in protein-folding chaperone recruitment would be critical during punctuated bursts of venom production post envenomation, as the venom store in the gland is rapidly replenished. Similarly, transcription factors involved in epithelial development that show increased expression in the venom gland are undoutebly linked to demands to maintain the venom gland lumen during venom production. Collectively, our findings raise the possibility that a core set of venom gland-specific transcription factors function to co-regulate venom production in venom gene clusters of open chromatin, and illustrate that venom production may be made possible through increased activity of other transcription factors involved in cellular stress responses and development.

Conclusion

Our analysis of the prairie rattlesnake genome provides new, and in some cases, surprising insight into the structure and function of reptilian and snake genomes, and broadly argues for the importance of studying diverse vertebrate lineages to understand the scope of vertebrate genome structure and function. For example, it appears that snakes have re-evolved genomic isochore structure not through an accumulation of GC content as observed in mammals and birds, but rather through the accumulation of AT content, suggesting a distinct GC isochore generative mechanism in snake genomes. Evidence for distinct evolutionary strata and the pseudoautosomal region of a snake sex chromosome, which bear unique hallmarks of the evolutionary trajectory from an ancestral autosomal chromosome pair, provide key comparative evidence to explain mechanisms underlying at least a majority of the partial dosage compensation observed in snakes. As the first species with microchromosomes to be analyzed at the nuclear organizational level using Hi-C, we show the surprising degree to which rattlesnake microchromosomes physically contact and interact with other chromosomes in the nucleus, suggesting that microchromosomes may operate in a fundamentally different way than macrochromosomes. Finally, in addition to the medical importance of studying snake venom, snake venom systems represents an intriguing model for understanding how evolution can direct the organization and regulation of a novel organ system – the venom gland – one of nature's most dynamic trophic adaptations. The excellent contiguity of our genome assembly enabled the definitive chromosomal localization of venom gene clusters, most of which are found on microchromosomes, and illustrates clearly the mechanistic process of tandem duplication that has given rise to venom gene diversity multiple times across venom gene families. Our results

also demonstrate many new and exciting mechanisms that underlie the tight regulation of venom genes, and the coordinated roles of chromatin and specific transcription factors in this process, as well as the co-evolution of other cellular mechanisms required to meet the extreme demands of bursts of venom production. Despite the key perspectives that the rattlesnake genome provides, many open questions remain, such as the evolutionary mechanisms by which snakes have accumulated AT content, how venom genes have gained venom gland-specific transcription factor binding sites, and the degree to which chromatin state is modulated in other tissues to prevent toxic venom gene expression. Conclusions from this and other studies consistently point to the unique and extreme biology of snakes that also extends to the unique biology of their genomes, highlighting the value of snakes and other non-traditional models in delivering new and often surprising perspectives into vertebrate biology and evolution.

Figures

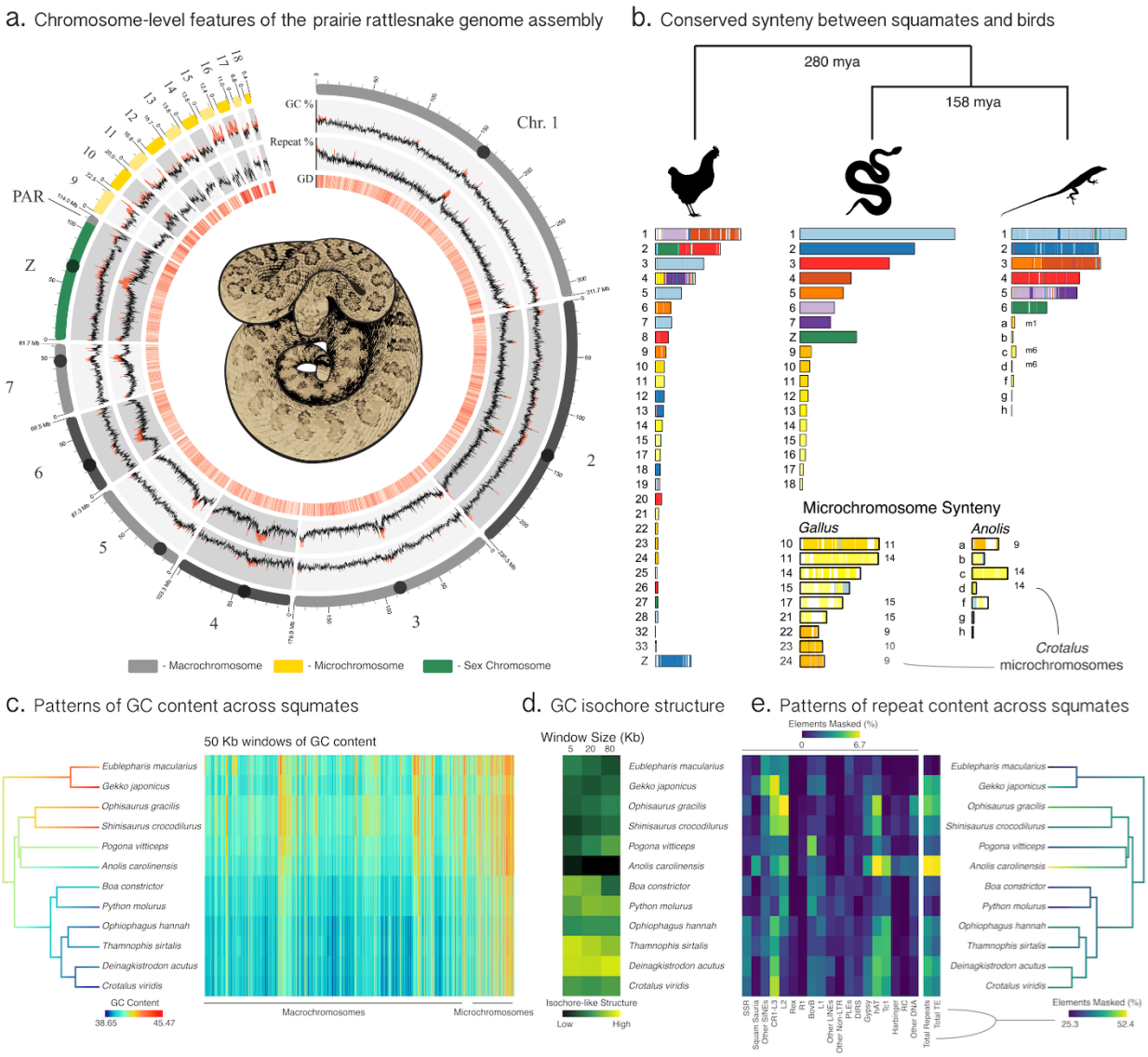


Figure 1. The chromosome-level prairie rattlesnake genome assembly. a, Diagram of genome-wide statistics. Chromosomes are show to scale and tick marks represent megabases of sequence. Chromosome ideogram is shown in the outer band, where circles represent inferred centromere locations. The grey segment of the Z chromosome represents the candidate pseudoautosomal region. GC% is the proportion of GC in 100Kb windows and Repeat % is the proportion of bases annotated as repeat content within each 100Kb window. Values above the genome-wide median are in red. GD is gene density, or the number of genes per 100Kb window; higher density is represented by darker red bands. **b**, Synteny between the rattlesnake and chicken and anole lizard genomes. Colors on chicken and anole chromosomes correspond with homologous sequence in rattlesnake. In the microchromosome inset, numbers to the right of chromosomes represent rattlesnake microchromosomes with which a given chicken or anole chromosome was syntenic for greater than 80% of its length. **c,** Tree branches are colored according to genomic GC content. The heatmap to the right depicts GC content in consecutive 50 Kb windows from a whole genome alignment, with macro- and microchromosome regions delineated. **d,** Genomic GC isochore structure measured by the standard deviation in GC content among 5, 20, and 80 Kb windows. **e,** Repeat content among 12 squamate species. Tree branches are colored by genomic repeat content.

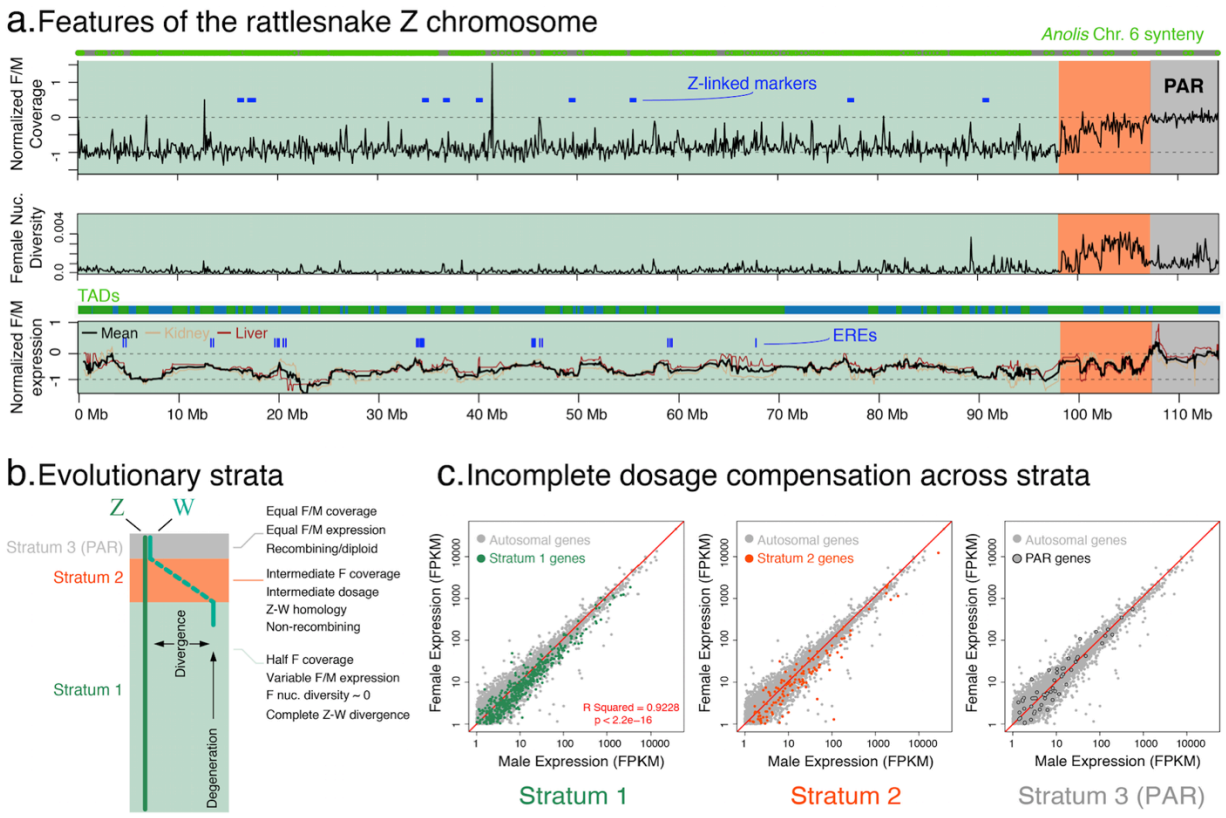


Figure 2. The Z chromosome of the prairie rattlesnake. a, Chromosomal landscape of log2 normalized female/male coverage, female nucleotide diversity, and log2 normalized female/male gene expression. High similarity BLAST hits to *Anolis* chromosome 6 are shown at the top as grey and green circles (high-stringency hits are in green; see Supplementary Methods). The positions of Z-linked cDNA markers from Matsubara et al. (2006) are shown as blue blocks, and the intervals of chromatin domains (TADs) are depicted as alternating green and blue blocks above the normalized expression plot. In the normalized expression plot, blue vertical lines represent the positions of predicted estrogen response elements (EREs) within 100 Kb of a dosed gene. On each plot, the pseudoautosomal region (PAR) and evolutionary Stratum 2 are highlighted in grey and orange, respectively, and Stratum 3 is highlighted in green. **b,** Schematic of the hypothesized evolutionary strata on the rattlesnake Z chromosome, with features that define them to the right. The dashed blue-green line representing the W chromosome depicts the inferred intermediate level of divergence between the Z and W chromosomes along Stratum 2. **c,** Patterns of relative female and male gene expression in each evolutionary stratum, plotted against the autosomal background (grey).

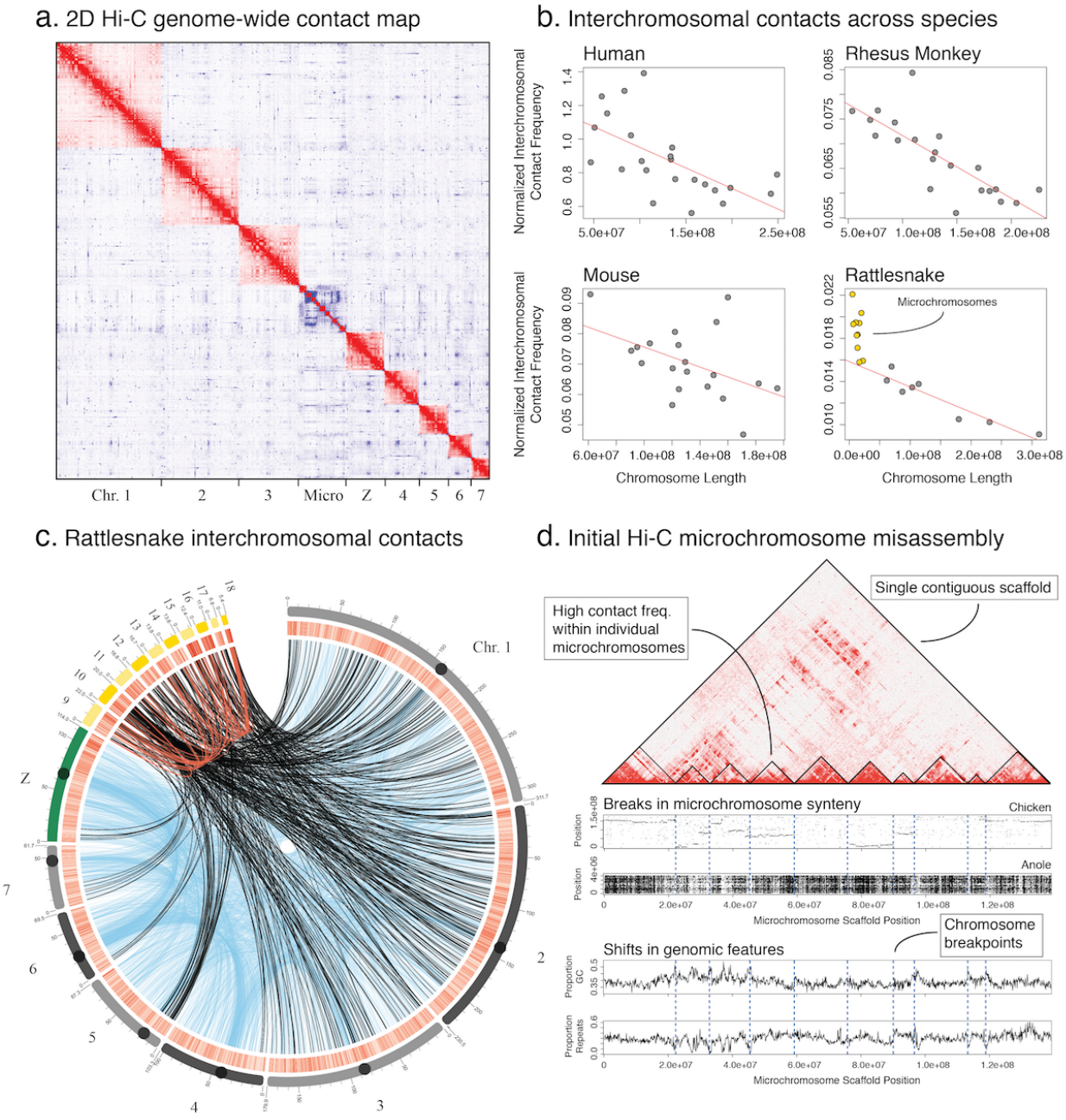


Figure 3. Genome-wide chromosomal contacts in the rattlesnake venom gland. a, 2D heatmap of intrachromosomal (red) and interchromosomal (blue) contacts among rattlesnake chromosomes. Higher color intensity depicts higher contact frequency. **b,** Comparison of interchromosomal contacts, normalized by chromosome length, and chromosome length between mammalian Hi-C datasets and the rattlesnake. Red lines depict the negative linear relationship between interchromosomal contacts and chromosome length for macrochromosomes. **c,** Locations of high-frequency interchromosomal contacts among rattlesnake chromosomes. Blue lines represent inter-macrochromosome contacts, black lines represent micro-to-macrochromosome contacts, and red lines represent inter-microchromosome contacts. **d,** Schematic of the initial misassembled microchromosome scaffold. The heatmap panel at the top depicts the high frequency inter- and intrachromosomal contacts among microchromosomes, and black triangles depict boundaries between microchromosomes. The middle two panels show synteny alignments between rattlesnake, chicken, and anole microchromosomes. The bottom two panels show windowed GC and repeat content across microchromosomes. Blue dashed lines in the lower panels show breakpoints between individual microchromosomes.

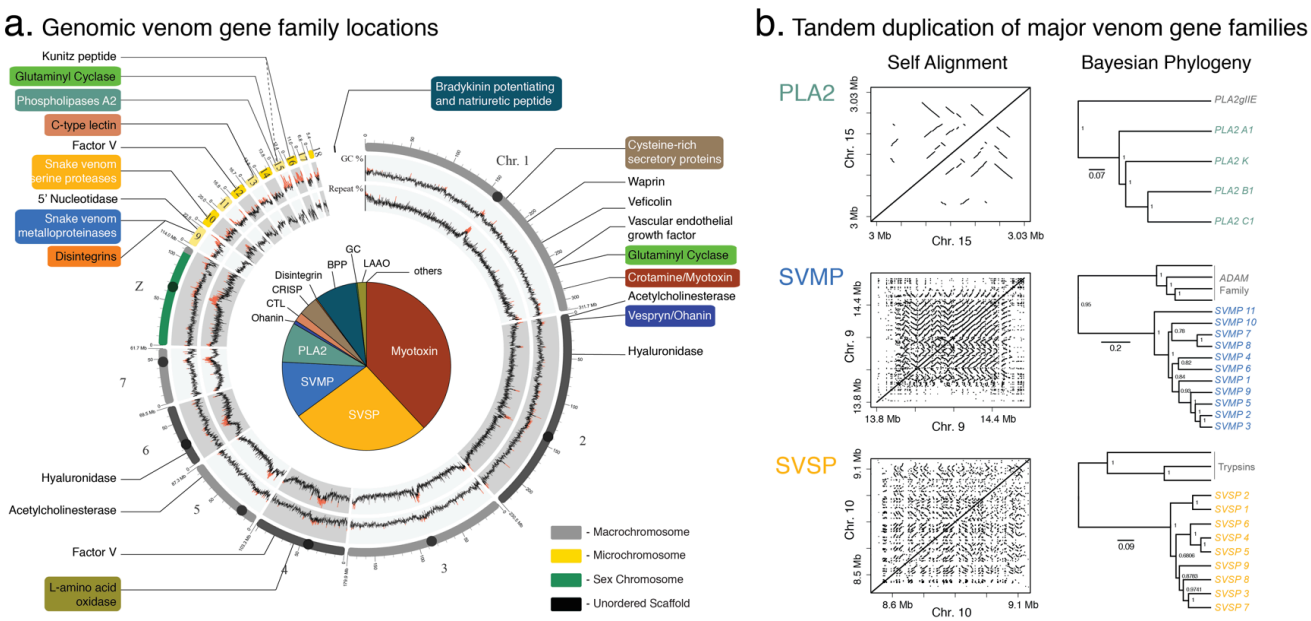


Figure 4. Genomic location of venom gene families and evidence for venom gene evolution through tandem duplication. **a,** The pie chart on the inside of the circularized genome ideogram represents the prairie rattlesnake proteome, redrawn from Saviola et al. 2015. The genome ideogram, GC content, repeat content, and legend at the bottom right follow the description in figure 1. Outside labels point to the genomic location of each venom gene family. **b,** Regional self alignment of phospholipase A2 (PLA2), snake venom metalloproteinase (SVMP), and serine proteinase (SVSP) venom gene clusters (left). Parallel and perpendicular lines off of the central diagonal line indicate segmental duplications. Bayesian phylogenetic tree estimates for each of the three gene families (right). Values at nodes represent posterior probabilities.

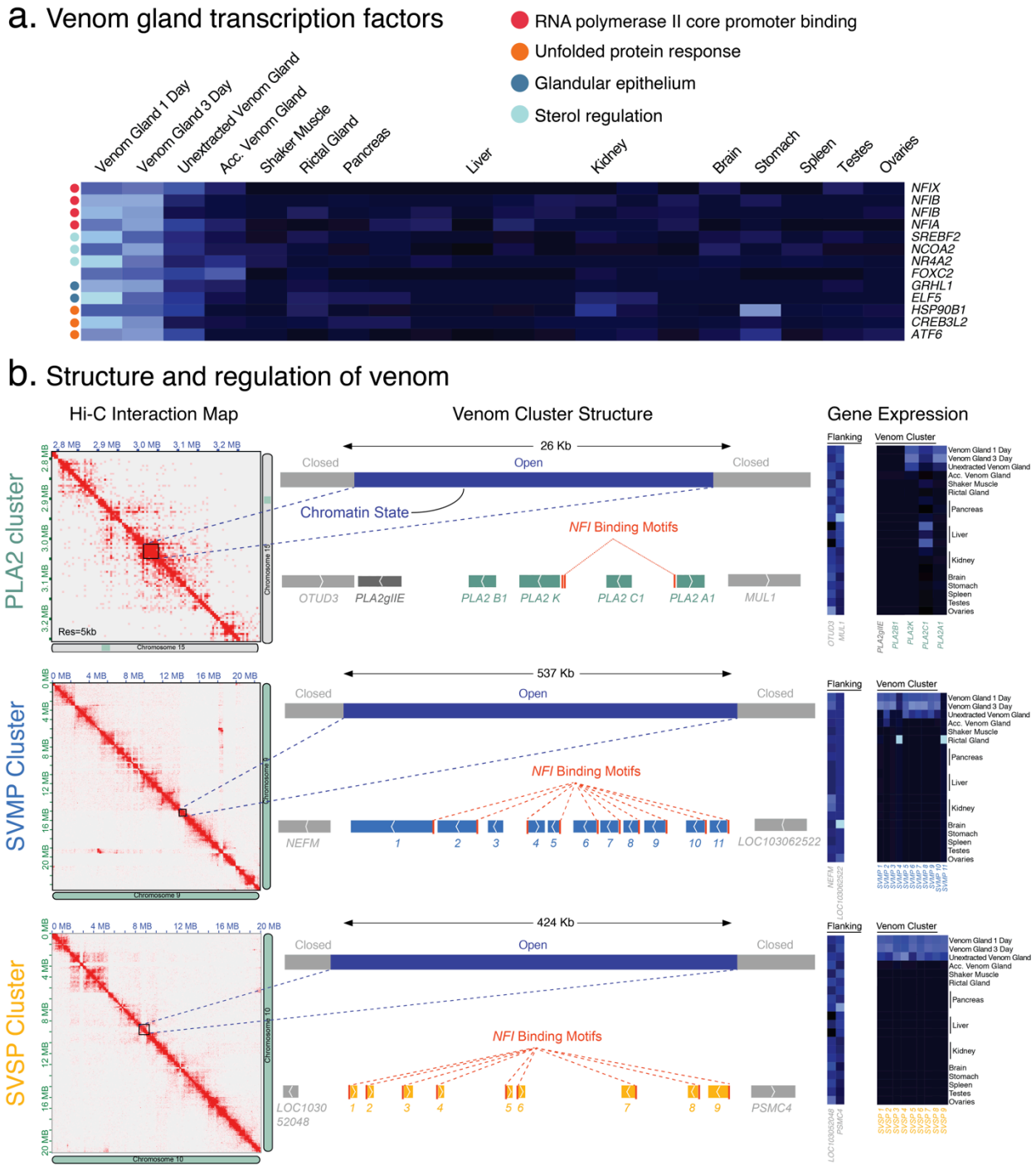


Figure 5. Mechanisms of venom gene regulation. a, Gene expression of transcription factors found to be significantly upregulated in venom glands, with expression values shown across tissues. Brighter colors show higher gene expression. The colored dots to the left of heatmap rows correspond to UniProt classifications of each transcription factor, which generally fell into the four categories in the legend in the top right. **b,** Genomic structure and regulation of the PLA2, SVMP, and SVSP venom gene families. 2D Hi-C contact maps are shown to the left, and boxes are used to show the bounds of each venom gene region. The schematics in the center depict the inferred chromatin state of each venom gene region (i.e., open chromatin) in the venom gland, the structure of each venom gene family and the non-venom genes flanking them. Predicted *NFI* transcription factor binding sites are shown as orange boxes upstream of genes. Gene expression profiles are shown to the right for each venom gene family and the flanking genes.

Supplementary Methods

Prairie rattlesnake Genome Sequencing and Assembly

A male prairie rattlesnake (*Crotalus viridis viridis*) collected from a wild population in Colorado was used to generate the genome sequence. This specimen was collected and humanely euthanized according to University of Northern Colorado Institutional Animal Care and Use Committee protocols 0901C-SM-MLChick-12 and 1302D-SM-S-16. Colorado Parks and Wildlife scientific collecting license 12HP974 issued to S.P. Mackessy authorized collection of the animal. Genomic DNA was extracted using a standard Phenol-Chloroform-Isoamyl alcohol extraction from liver tissue that was snap frozen in liquid nitrogen. Multiple short-read sequencing libraries were prepared and sequenced on various platforms, including 50bp single-end and 150bp paired-end reads on an Illumina GAII, 100bp paired-end reads on an Illumina HiSeq, and 300bp paired-end reads on an Illumina MiSeq. Long insert libraries were also constructed by and sequenced on the PacBio platform. Finally, we constructed two sets of mate-pair libraries using an Illumina Nextera Mate Pair kit, with insert sizes of 3-5Kb and 6-8Kb, respectively. These were sequenced on two Illumina HiSeq lanes with 150bp paired-end sequencing reads. Short and long read data were used to assemble the previous genome assembly version CroVir2.0 (NCBI accession SAMN07738522). Details of these sequencing libraries are in Supplementary Table 1. Prior to assembly, reads were adapter trimmed using BBmap (Bushnell 2014) and we quality trimmed all reads using Trimmomatic v0.32 (Bolger et al. 2014). We used Meraculous (Chapman et al. 2011) and all short-read Illumina data to generate a contig assembly of the prairie rattlesnake. We then performed a series of scaffolding and gap-filling steps. First, we used L_RNA_scaffolder (Xue et al. 2013) to scaffold contigs using the complete

transcriptome assembly (see below), SSPACE Standard (Boetzer et al. 2010) to scaffold contigs using mate-pair reads, and SSPACE Longread to scaffold using long PacBio reads. We then used GapFiller (Nadalin et al. 2012) to extend contigs and fill gaps using all short-read data cross five iterations. We merged the scaffolded assembly with a contig assembly generated using the *de novo* assembly tool in CLC Genomics Workbench (Qiagen Bioinformatics, Redwood City, CA, USA).

We improved the CroVir2.0 assembly using the Dovetail Genomics HiRise assembly method, leveraging both Chicago and Hi-C sequencing. Chicago assembly requires large amounts of high molecular weight DNA from a very fresh tissue sample. We thus extracted high molecular weight genomic DNA from a liver of a closely related male to the CroVir2.0 animal (i.e., from the same den site). This animal was collected and humanely euthanized according to the Colorado Parks and Wildlife collecting license and UNC IACUC protocols detailed above. Hi-C sequencing data were derived from the venom gland of the same animal (see details below on venom gland Hi-C and RNAseq experimental design). Dovetail Genomics HiRise assembly resulted in a highly contiguous genome assembly (CroVir3.0) with a physical coverage of greater than 1,000x. We estimated the size of the genome using k-mer frequency distributions (17, 19, and 21mers) quantified using Jellyfish (Marçais and Kingsford 2011).

We generated transcriptomic libraries from RNA sequenced from 16 different tissues: two venom gland tissues; 1 day and 3 days post-venom extraction (see Hi-C and RNA sequencing of Venom Gland section below), one from pancreas, and one from tongue were taken from the Hi-C sequenced genome animal. Additional samples from other individuals included a third venom

gland sample from which venom had not been extracted (‘unextracted venom gland’), three liver, three kidney, two pancreas, and one each of skin, lung, testis, accessory venom gland, shaker muscle, brain, stomach, ovaries, rictal gland, spleen, and blood tissues. Total RNA was extracted using Trizol, and we prepared RNAseq libraries using an NEB RNAseq kit for each tissue, which were uniquely indexed and run on multiple HiSeq 2500 lanes using 100bp paired-end reads (Supplementary Table 6). We used Trinity v. 20140717 (Grabherr et al. 2011) with default settings and the ‘--trimmomatic' setting to assemble transcriptome reads from all tissues. The resulting assembly contained 801,342 transcripts comprising 677,921 Trinity-annotated genes, with an average length of 559 bp and an N50 length of 718 bp.

Repeat Element Analysis

Annotation of repeat elements was performed using homology-based and *de novo* prediction approaches. Homology-based methods of transposable element identification like *RepeatMasker* cannot recognize elements that are not in a reference database, and have low power to identify fragments of repeat elements belonging to even moderately diverged repeat families (Platt et al. 2016). Since the current release of the Tetrapoda RepBase library (Bao et al. 2015) (v.20.11, August 2015) is unsuitable for detailed repeat element analyses of most squamate reptile genomes, we performed *de novo* identification of repeat elements on 6 snake genomes (*Crotalus viridis*, *Crotalus mitchellii*, *Thamnophis sirtalis*, *Boa constrictor*, *Deinagkistrodon acutus*, and *Pantherophis guttatus*) in RepeatModeler v.1.0.9 (Smit and Hubley 2015) using default parameters. Consensus repeat sequences from multiple species were combined into a large joint snake repeat library that also includes previously identified elements from an additional 12 snake species (Castoe et al. 2013). All genomes were annotated with the same library with the

exception of the green anole lizard, for which we used a lizard specific library that includes *de novo* repeat identification for *Pogona vitticeps*, *Ophisaurus gracilis*, and *Gekko japonicus*. To verify that only repeat elements were included in the custom reference library, all sequences were used as input in a BLASTx search against the SwissProt database (UniProt 2017), and those clearly annotated as protein domains were removed. Finally, redundancy and possible chimeric artifacts were removed through clustering methods in CD-HIT (Li and Godzik 2006) using a threshold of 0.85.

Homology-based repeat element annotation was performed in RepeatMasker v.4.0.6 (Smit et al. 2015) using a PCR-validated BovB/CR1 LINE retrotransposon consensus library (Castoe et al. 2013), the Tetrapoda RepBase library, and our custom library as references. Output files were post-processed using a modified implementation of the ProcessRepeat script (RepeatMasker package).

Gene Annotation

We used MAKER v. 2.31.8 (Cantarel et al. 2008) to annotate protein-coding genes in an iterative fashion. Several sources of empirical evidence of protein-coding genes were used, including the full *de novo C. viridis* transcriptome assembly and protein datasets consisting of all annotated proteins from NCBI for *Anolis carolinensis* (Alfoldi et al. 2011), *Python molurus bivittatus* (Castoe et al. 2013), *Thamnophis sirtalis* (Perry et al. In Review), and *Ophiophagus hannah* (Vonk et al. 2013), and from GigaDB for *Deinagkistrodon acutus* (Yin et al. 2016). We also included 422 protein sequences for 24 known venom gene families that were used to infer *Python* venom gene homologs in a previous study (Reyes-Velasco et al. 2015). Prior to running

MAKER, we used BUSCO v. 2.0.1 (Simão et al. 2015) and the full *C. viridis* genome assembly to iterative train AUGUSTUS v. 3.2.3 (Stanke and Morgenstern 2005) HMM models based on 3,950 tetrapod vertebrate benchmarking universal single-copy orthologs (BUSCOs). We ran BUSCO in the 'genome' mode and specified the '--long' option to have BUSCO perform internal AUGUSTUS training. We ran MAKER with the 'est2genome=0' and 'protein2genome=0' options set to produce gene models using the AUGUSTUS gene predictions with hints supplied from the empirical transcript and protein sequence evidence. We provided the coordinates for all interspersed, complex repetitive elements for MAKER to perform hard masking before evidence mapping and prediction, and we set the 'model_org' option to 'simple' to have MAKER soft mask simple repetitive elements. We used default settings for all other options, except 'max_dna_len' (set to 300,000) and 'split_hit' (set to 20,000). We iterated this approach an additional time and we manually compared the MAKER gene models with the transcript and protein evidence. We found very little difference between the two gene annotations and based on a slightly better annotation edit distance (AED) distribution in the first round of MAKER, we used our initial round as the final gene annotation. The resulting annotation consisted of 17,486 genes and we ascribed gene IDs based on homology using reciprocal best-blast (with e-value thresholds of 1e-5) and stringent one-way blast (with an e-value threshold of 1e-8) searches against protein sequences from NCBI for *Anolis*, *Python*, and *Thamnophis*.

Hi-C and RNA Sequencing of the Venom Gland

We dissected the venom glands from the Hi-C *Crotalus viridis viridis* 1 day and 3 days after venom was initially extracted in order to track a time-series of venom production. A subsample of the 1-day venom gland was sent to Dovetail Genomics where DNA was extracted and

replicate Hi-C sequencing libraries were prepared according to their protocol (see above). We also extracted total RNA from both 1-day and 3-day venom gland samples, as well as tongue and pancreas tissue from the Hi-C genome animal (see Sequencing and Assembly and Annotation sections above). mRNAseq libraries were generated and sequenced at Novogene on two separate lanes of the Illumina HiSeq 4000 platform using 150 bp paired-end reads (Supplementary table 6).

Chromosome Identification and Synteny Analyses

Genome assembly resulted in several large, highly-contiguous scaffolds with a relative size distribution consistent with the karyotype for *C. viridis* (Baker et al. 1972), representing nearly-complete chromosome sequences. We determined the identity of chromosomes using a BLAST search of the chromosome-specific markers linked to snake chromosomes from (Matsubara et al. 2006), downloaded from NCBI (accessions SAMN00177542 and SAMN00152474). We kept the best alignment per cDNA marker as its genomic location in the Prairie Rattlesnake genome, except when a marker hit two high-similarity matches on different chromosomes. The vast majority of markers linked to a specific macrochromosome (i.e., chromosomes 1-7; Supplementary Table 4) in *Elaphe quadrivirgata* mapped to a single genomic scaffold; only 6 of 104 markers did not map to the predicted chromosome from *E. quadrivirgata*. All snake microchromosome markers mapped to a single 139Mb scaffold, which was later broken into 10 microchromosome scaffolds (scaffold-mi1-10; see below).

We identified a single 114Mb scaffold corresponding to the Z chromosome, as 10 of 11 Z-linked markers mapped to this scaffold. To further vet this as the Z-linked region of the genome, we

mapped reads from male and female *C. viridis* (Supplementary Table 7) to the genome using BWA (Li and Durbin 2009) using program defaults. Male and female resequencing libraries were prepared using an Illumina Nextera prep kit and sequenced on an Illumina HiSeq 2500 using 250bp paired-end reads. Adapters were trimmed and low-quality reads were filtered using Trimmomatic (Bolger et al. 2014). After mapping, we filtered reads with low mapping scores and quantified per-base read depths using SAMtools (Li et al. 2009). We then binned read depths into 100Kb windows and normalized female and male windowed-coverage by calculating the log2(female/male) ratio. Here, the expectation is that a hemizygous locus will show roughly half the normalized coverage, which we observe for females over the majority of the Z chromosome scaffold length, and not elsewhere in the genome (Supplementary Fig. 4). To demonstrate Z chromosome conservation among pit vipers and to further determine the identity of this scaffold, we mapped male and female pygmy rattlesnake (*Sistrurus catenatus*) reads from (Vicoso et al. 2013) to the genome using the same parameters detailed above (Supplementary Fig. 4). *Anolis* chromosome 6 is homologous with snake sex chromsomes (Srikulnath et al. 2009), thus we aligned *Anolis* Chromosome 6 (Alfoldi et al. 2011) to the Prairie Rattlesnake genome using BLASTn. As expected, we found a large quantity of high-similarity hits to the rattlesnake Z chromosome scaffold, specifically, which were organized in a sequential manner across the entire Z scaffold (Figs. 1b, 2a). In Fig. 2a, 'high-stringency' hits refer to alignments with e-values $< 1e^{-250}$ and bit-scores > 200. The gray circles, which appear as a solid line due to their density, correspond with hits with an e-value < 0.00001.

We used multiple sources of information to identify the best candidate breakpoints between microchromosomes within the 139Mb fused microchromosome scaffold in the initial Hi-C

assembly. Because reptile microchromosomes are highly syntenic (Alfoldi et al. 2011), we aligned the microchromosome scaffold to microchromosome scaffolds from chicken (Hillier et al. 2004) and *Anolis* using LASTZ (Harris 2007) to determine likely chromosomal breakpoints. To retain only highly similar alignments per comparison, we set the 'hspthresh' option equal to 10,000 (default is 3,000). We also set a step size equal to 20 to reduce computational time per comparison. This approach delineated candidate boundaries between rattlesnake microchromosomes based on clear breaks in cross-species synteny (Fig. 3d). We further validated candidate break points using genomic features that consistently vary at the ends of chromosomes. Here, we specifically evaluated if candidate breakpoints exhibited regional shifts in GC content and repeat content, similar to the ends of macrochromosomes (Fig. 1). For each candidate breakpoint, we then determined if there was a junction between two Chicago assembly scaffolds (i.e., two contiguous pieces of sequence that were not assembled using Hi-C) within the breakpoint region. Finally, if no annotated genes spanned this junction, we considered it biologically plausible. There were nine candidate breakpoints that met each of these criteria, equaling the number of boundaries expected given ten microchromosomes (Fig. 3d). Importantly, this approach assumes that the ten microchromosomes were assembled in a contiguous fashion per chromosome. Intrachromosomal chromatin contacts are far more frequent than contacts between chromosomes (Lieberman-Aiden et al. 2009). The ten candidate microchromosomes match this expectation, and show clear signal of consistent intrachromosomal contact frequencies across their entire length (the same as macrochromosomes; Supplementary Fig. 9).

To explore broad-scale structural evolution across reptiles, we used the rattlesnake genome to perform in silico painting of the chicken (*Gallus gallus* version 5) and *Anolis carolinensis* (version 2) genomes. Briefly, we divided the rattlesnake genome into 2.02 million potential 100 bp markers. For each of these markers, we used BLAST to record the single best hit in the target genome requiring an alignment length of at least 50 bp. This resulted in 41,644 potential markers in *Gallus* and 103,801 potential markers in *Anolis*. We then processed markers on each chromosome by requiring at least five consecutive markers supporting homology to the same rattlesnake chromosome. We consolidated each group of five consecutive potential markers as one confirmed marker. In *Gallus*, we rejected 12.4% of potential markers and identified 7,291 confirmed merged markers. In *Anolis*, we rejected 39.7% of potential markers and identified 12,511 confirmed merged markers.

This approach demonstrates considerable stability at the chromosomal level despite 158 million years of divergence between *Anolis* and *Crotalus* (Fig. 1b), and between squamates and birds, despite 280 million years of divergence between squamates and *Gallus*. This stability is evident not only in the macrochromosomes but also in the microchromosomes. In fact, 7 of 10 *Crotalus* microchromosomes had greater than 80% of confirmed markers associated with a single chromosome in the chicken genome (Fig. 1b, microchromosome inset). Comparisons among the three genomes suggest that the *Crotalus* genome has not experienced some of the fusions found in *Anolis*. Specifically, we infer that *Anolis* chromosome 3 is a fusion of *Crotalus* chromosome 4 and 5. Likewise, *Anolis* chromosome 4 is a fusion of *Crotalus* chromosome 6 and 7. Divergence time estimates discussed above and shown in Fig. 1b were taken from the median of estimates

for divergence between *Crotalus* and *Gallus* and between *Crotalus* and *Anolis* from Timetree (www.timetree.org; (Kumar et al. 2017).

Genomic Patterns of GC Content

We quantified GC content in sliding windows of 100Kb and 1Mb across the genome using a custom Python script (https://github.com/drewschield/Comparative-Genomics-Tools/blob/master/slidingwindow_gc_content.py). GC content in 100Kb windows is presented in Fig. 1 in the Main Text.

To determine if there is regional variation in nucleotide composition consistent with isochore structures across the rattlesnake genome, we quantified GC content and its variance within 5, 10, 20, 40, 80, 160, 240, and 320-kb windows. The variation (standard deviation) in GC content is expected to decrease by half as window size increases four-fold if the genome is homogeneous (i.e., lacks isochore structures; (Consortium 2001). By comparing the observed variances of GC content across spatial window scales to those from 11 other squamate genomes, including lizards (*Anolis* has been shown to lack isochore structure (Alfoldi et al. 2011), henophidian snakes, and colubroid snakes, we were able to determine the relative heterogeneity of nucleotide composition in the rattlesnake (Supplementary Table 8). To reduce potential biases from estimates from small scaffold sizes, we filtered to only retain scaffolds greater than the size of the window analyzed (e.g., only scaffolds longer than 10 Kb when looking at the standard deviation in GC content over 10 Kb windows) and for which more there was less than 20% of missing data for all analyzed genomes.

To explore trajectories of GC content evolution among squamates, we generated whole genome alignments for the species in Supplementary Table 8 using Multi-Z (Blanchette et al. 2004), using program defaults. We then filtered the multi-species whole genome alignment to retain only blocks for which information for all 12 species was available, and concatenated blocks according to their organization in the *Anolis* lizard genome. We then calculated GC content within consecutive 50 Kb windows of this concatenated alignment using the 'slidingwindow_gc_content.py' script detailed above.

Hi-C analysis

Raw Illumina paired-end reads were processed using the Juicer pipeline (Durand et al. 2016) to produce Hi-C maps binned at multiple resolutions, as low as 5kb resolution, and for the annotation of contact domains. These data were aligned against the CroVir3.0 assembly. All contact matrices used for further analysis were KR-normalized in Juicer. We identified topologically-associated chromatin domains (TADs) using the Hi-C Explorer 'hicFindTADs' function (Ramírez et al. 2018), using default parameters and specifying a Bonferroni correction for multiple comparisons.

We compared intra and interchromosomal contact frequencies between the rattlesnake venom gland and various tissues from mammals. To do this we quantified the total intra- and interchromosomal contacts between chromosome positions from the rattlesnake and the following Hi-C datasets: human lymphoblastoma cells (Rao et al. 2014) and human retinal ephithelial cells, mouse kidney, and Rhesus monkey tissue (Darrow et al. 2016). To investigate patterns of intra- and interchromosome contact frequency, we normalized contact frequencies by

chromosome length. In the case of the mouse, we removed the Y chromosome due to its small size and relative lack of interchromosomal contacts. We then performed linear regressions of chromosome length and normalized intra- and interchromosomal contact frequencies (i.e., contact frequency/chromosome length). In all cases we observed a positive relationship between normalized intrachromosomal contacts and chromosome size and a negative relationship between normalized interchromosomal contacts and chromosome size (Fig. 3b).

Sex Chromosome Analysis

We identified the Prairie Rattlesnake Z chromosome using methods described in section 1.X above. We localized the candidate pseudoautosomal region (PAR) based on normalized female/male coverage (Fig. 2a; the PAR is the only consistent region of the Z with equal female and male coverage. We quantified gene content, GC content, and repeat content across the Z chromosome and PAR (Supplementary Figs. 5, 6, and 7), and tested for gene enrichment in the PAR using a Fisher's exact test, where we compared the number of genes within each region to the total length of the region.

To compare within individual nucleotide diversity across the genome between male and female *C. viridis*, we called variants (i.e., heterozygous sites) from the male and female reads used in coverage analysis detailed above. With the mappings from coverage analysis, we used SAMtools (Li et al. 2009) to compile all mappings into pileup format, from which we called variant sites using BCFtools. We filtered sites to retain only biallelic variants using VCFtools (Danecek et al. 2011) and calculated the proportion of heterozygous sites (i.e., within-individual nucleotide diversity) using a custom pipeline of scripts. First, calcHet

(https://github.com/darencard/RADpipe) outputs details of heterozygous site and window_heterozygosity.py (https://github.com/drewschield/Comparative-Genomics-Tools/blob/master/window_heterozygosity.py) uses this output in conjunction with a windowed .bed file generated using BEDtools 'make_windows' tool to calculate the proportion of heterozygosity within a given window size.

Evolutionary patterns of the Z chromosome were also analyzed by examining transposable element age and composition along the whole chromosome, and across the three inferred evolutionary strata (see Main Text). Since the length of the PAR is significantly smaller than the combined length of Strata 1 and 2, to rule out potential biases due to unequal sample size we also independently analyzed fragments of the other strata with lengths equal to the PAR (total of 15 7.18 Mbp fragments). Each region was analyzed in RepeatMasker using a single reference library that included the squamate fraction of the RepBase Tetrapoda library, and the snake specific library clustered at a threshold of 0.75. The age distribution of TE families was estimated by mean of the Kimura 2-parameter distance from the consensus sequence per element (CpG corrected) calculated from PostProcessed.align outputs (see section 1.X above). We then merged estimates of repeat content from each of these regions for comparison to the PAR region, specifically.

To quantify gene expression on the rattlesnake Z chromosome and across the genome, we prepared RNAseq libraries from liver and kidney tissue from two males and females and sequenced them on an Illumina HiSeq using 100bp paired-end reads (Supplementary Table 6). Samples and libraries were prepared following the methods of (Andrew et al. 2017). After

filtering and adapter trimming using Trimmomatic v. 0.32 (Bolger et al. 2014), we mapped RNAseq reads to the *C. viridis* genome using STAR v. 2.5.2b (Dobin et al. 2013) and counts were determined using featureCounts (Liao et al. 2013). We normalized read counts across tissues and samples using TMM normalization in edgeR (Robinson et al. 2010) to generate both counts per million (CPM) for use in pairwise comparisons between males and females, and fragments per kilobase million (FPKM) normalized counts for comparisons of chromosome-wide expression within samples. We tested for differential gene expression between males and females using pairwise exact tests in edgeR followed by independent hypothesis weighting (IHW) p-value correction (Ignatiadis et al. 2016) and quantified normalized gene expression across the Z chromosome in 100Kb windows, based on the location of each gene in the genome annotation. Per gene female-to-male ratios of normalized expression were generated by dividing the average female expression level by that of the male, only including genes with expression information in both the male and female (>1 avg. FPKM in each sex). Two-sided student's t-tests in R were used to compare of median female-to-male ratios between chromosomes and/or chromosomal regions (i.e. the PAR). To explore regional variation in dosage across the Z chromosome, we performed a sliding window analysis of the F/M log2 normalized expression ratio with a window size of 30 genes and a step size of 1 gene.

A possible mechanism for upregulation of certain Z-linked genes in females is regulation through estrogen response elements (EREs), which can enable binding of enhancers and promote transcription of genes over long distances (Lin et al. 2007). Rice et al. (2017) identified that the binding domain of *ESR1* is completely conserved among humans, chickens, and alligators, thus we used the *ESR1* binding motif of humans ('GGTCAnnnTGACC'; (Lin et al. 2007) and a

regular expression motif finding script (https://github.com/dariober/bioinformatics-cafe/tree/master/fastaRegexFinder) to predict *ESR1* binding motifs (ER motif) throughout the rattlesnake genome. Using BEDtools 'closest' function (Quinlan and Hall 2010), we calculated the distance from each gene to the nearest predicted ER motif. We considered a gene to be a candidate for ERE-based upregulation if it was within 100Kb of a predicted ERE. We calculated the number of genes with evidence of partial dosage (i.e., genes with a F/M expression ratio greater than the lower bound of the autosomal 95% quantile), and used a Fisher's Exact test to determine if 'dosed' genes were enriched for proximity to EREs, which was not significant.

Comparative Microchromosome Genomics

To understand evolutionary shifts in microchromosome composition among amniotes, we compared measures of gene density, GC content, and repeat content of macro- and microchromosomes between the rattlesnake, anole (Alfoldi et al. 2011), bearded dragon (Georges et al. 2015; Deakin et al. 2016), chicken (Hillier et al. 2004), and zebra finch (Warren et al. 2010) genomes. These species were chosen because their scaffolds are ordered into chromosomes and because their karyotypes contain microchromosomes. For each genome, we quantified the total number of genes per chromosome, total number of G+C bases, and total bases masked as repeats in RepeatMasker. We then normalized each measure by the total length of macrochromosome and microchromosome sequences in each genome, then calculated the ratio of microchromosome:macrochromosome proportions. We then used Fisher's Exact Tests determine if one chromosome set possessed a significantly greater proportion of each measure. We generated a phylogenetic tree (Supplementary Fig. 2) for the five species based on

divergence time estimates from TimeTree (Kumar et al. 2017), and plotted the ratio values calculated above onto the tree tips for between-species comparisons.

Venom Gene Annotation and Analysis

We took a multi-step approach toward identifying venom gene homologs in the rattlesnake genome. We first obtained representative gene sequences for 38 venom gene families from Genbank (Supplementary Table 9), comprising known enzymatic and toxin components of snake venoms. We then searched our transcript set using the venom gene family query set using a tBLASTx search, defining a similarity cutoff e-value of $1x10^{-5}$. For each candidate venom gene transcript identified in this way, we then performed a secondary tBLASTx search against the NCBI database to confirm its identity as a venom gene. In the case of several venom gene families, such as those known only from elapid snake venom, we did not find any candidate genes. Three venom gene families that are especially abundant, both in terms of presence in the venom proteome (Fig. 4a) and in copy number, in the venom of *C. viridis* are phospholipases A2 (PLA2s), snake venom metalloproteinases (SVMPs), and snake venom serine proteases (SVSPs). Rattlesnakes possess multiple members of each of these families (Mackessy 2008; Casewell et al. 2011; Dowell et al. 2016), and the steps taken above appeared to underestimate the total number of copies in the *C. viridis* genome. Therefore, for each of these families, we performed an empirical annotation using the FGENESH+ (Solovyev et al. 2006) protein similarity search. We first extracted the genomic region annotated for each of these families above plus and minus a 100 Kb flanking region. We used protein sequences from Uniprot (PLA2: APD70899.1; SVMP: Q90282.1; and SVSP: F8S114.1) to query the region and confirm the total number of copies per family. Each gene annotated in this way was again searched against NCBI to confirm

its identity and manual searches of aligned protein sequences (see phylogenetic analyses below) further confirmed their homology to each respective venom gene family. Genomic locations and details of annotated venom genes in the rattlesnake genome are provided in Supplementary Table 10.

We used LASTZ (Harris 2007) to align the genomic regions containing PLA2, SVMP, and SVSP genes to themselves. We used program defaults, with the exception of the 'hspthresh' command, which we set to 8,000. This was done to only return very high similarity matches between compared sequences. Here the expectation is that when alignments are plotted against one another, we will observe a diagonal line demonstrating perfect matches between each stretch of sequence and itself. In the case of segmental duplications, we also expect to see parallel and perpendicular (if in reverse orientation) segments adjacent to the diagonal 'self' axis. We plotted LASTZ results for each of the regions using the base plotting function in R (R Core Team 2017).

We then performed Bayesian phylogenetic analyses to further evaluate evidence of tandem duplication and monophyly among members of the PLA2, SVMP, and SVSP venom gene families. We generated protein alignments of venom genes with their closest homologs using MUSCLE (Edgar 2004) with default parameters, with minor manual edits to the alignment to remove any poorly aligned regions. We analyzed the protein alignments using BEAST2 (Bouckaert et al. 2014), setting the site model to 'WAG' for each analysis. We ran each analysis for a minimum of 1×10^8 generations, and evaluated whether runs had reached stationarity using Tracer (Drummond and Rambaut 2007). After discarding the first 10% of samples as burnin, we

generated consensus maximum clade credibility trees using TreeAnnotator (distributed with BEAST2).

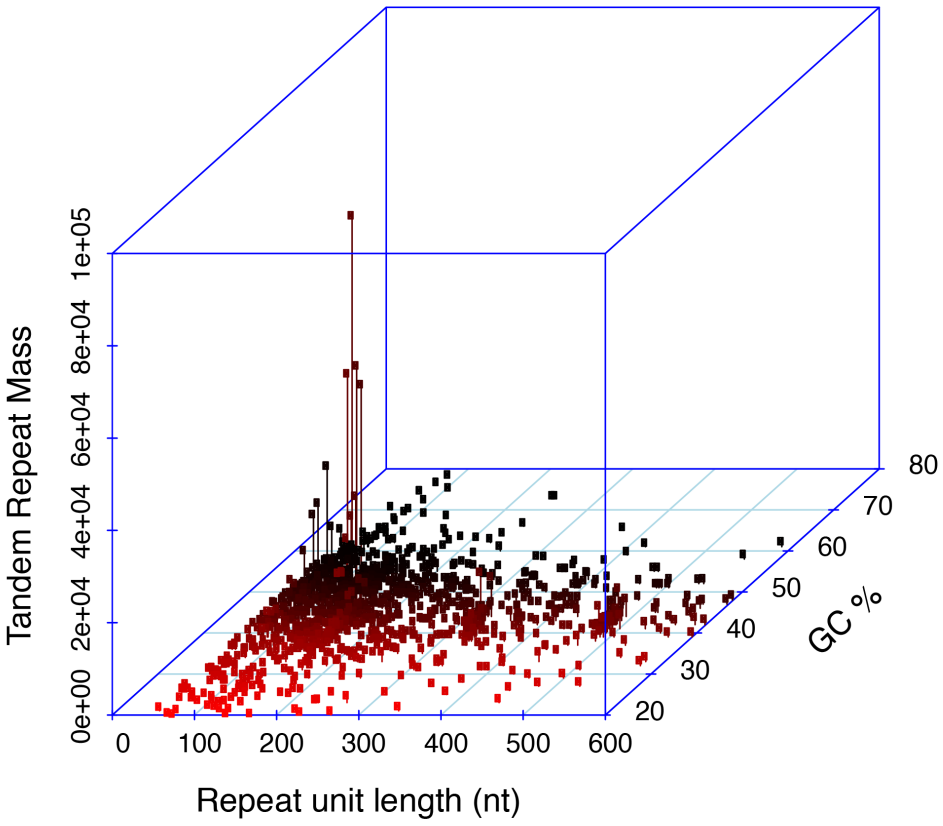
Raw Illumina RNAseq reads (Supplementary Table 6) were quality trimmed using Trimmomatic v. 0.36 (Bolger et al. 2014) with default settings. We used STAR (Dobin et al. 2013) to align reads to the genome. Raw expression counts were estimated by counting the number of reads that mapped uniquely to a particular annotated transcript using HTSeq-count (Anders et al. 2013). These raw counts were then normalized and filtered in edgeR using TMM normalization (Oshlack et al. 2010; Robinson et al. 2010), and all subsequent analyses were done using these normalized data. We used two-sided student's t-tests in R to compare gene expression between venom gland samples and body tissues to test for evidence of genes exhibiting a significantly upregulated signature of expression in the venom gland, specifically.

To identify candidate transcription factors regulating venom gene expression, we searched the genome annotation for all genes included on the UniProt (http://www.uniprot.org) reviewed human transcription factor database, by specifying species = 'Homo sapiens' and reviewed = 'yes' in the advanced search terms. Using this list, we parsed the rattlesnake genome for all matching gene IDs and compared their expression across rattlesnake tissues. We then identified likely candidate venom gland transcription factors, which showed a pattern of overall low body-wide expression and statistically significant evidence of higher expression in the venom gland, specifically. We found 13 candidates using this approach, including four members of the *CTF*/*NFI* family of RNA polymerase II core promoter-binding transcription factors (*NFIA,* two isoforms of *NFIB*, and *NFIX*). *NFI* binding sites have been identified upstream of venom genes

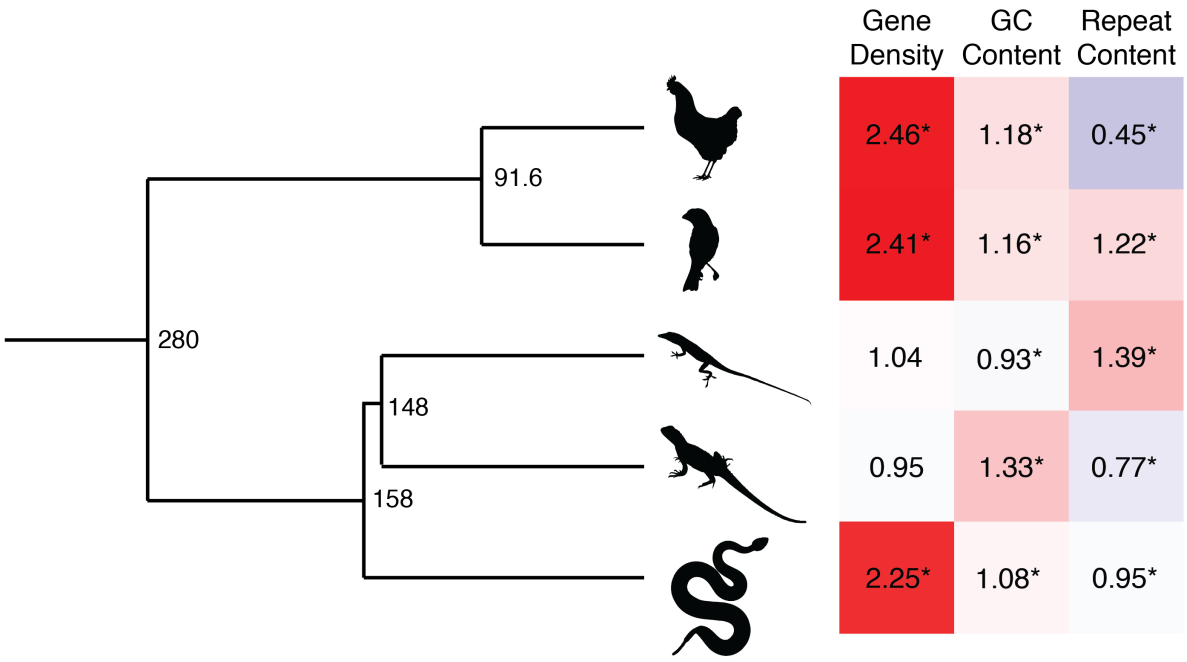
in several venomous snake taxa, including viperids, elapids, and colubrids (e.g., crotamine/myotoxin in *Crotalus durissus* (Rádis-Baptista et al. 2003) and three finger toxins in *Naja sputatrix* (Lachumanan et al. 1998) and *Boiga dendrophila* (Pawlak and Kini 2008). *NFI* family members were also found to be expressed in the venom glands of several species in a previous study exploring putative venom gland transcription factors (Hargreaves et al. 2014), but information about whether they showed venom gland-specific expression was not provided.

Because four transcription factors of the *NFI* family each showed evidence of venom gland-specificity, we tested the hypothesis that their binding motifs are also upstream of venom genes more than they are other genes. We obtained the TRANSFAC position weight matrix for each transcription factor from the CIS-BP database (Weirauch et al. 2014), scanned a 1 Kb region upstream of each gene in the snake venom PLA2, SVMP, and SVSP gene families for predicted transcription factor binding sites per upstream region using PoSSuM Search (Beckstette et al. 2006), setting a p-value cutoff of $1x10^{-6}$ for each search. We then performed the same analysis on 1 Kb regions of the closest related non-venom homologs per venom gene family, as well as the 1 Kb upstream regions of five independent random samples of 100 genes per sample. For each analyzed set of upstream regions, we performed a Fisher's Exact test of significant enrichment upstream of venom genes by comparing 1) the number of predicted binding motifs divided by the number of upstream regions, 2) the number of predicted binding motifs divided by the total combined length of upstream regions, and 3) the total length of predicted binding motifs divided by the total combined length of upstream regions.

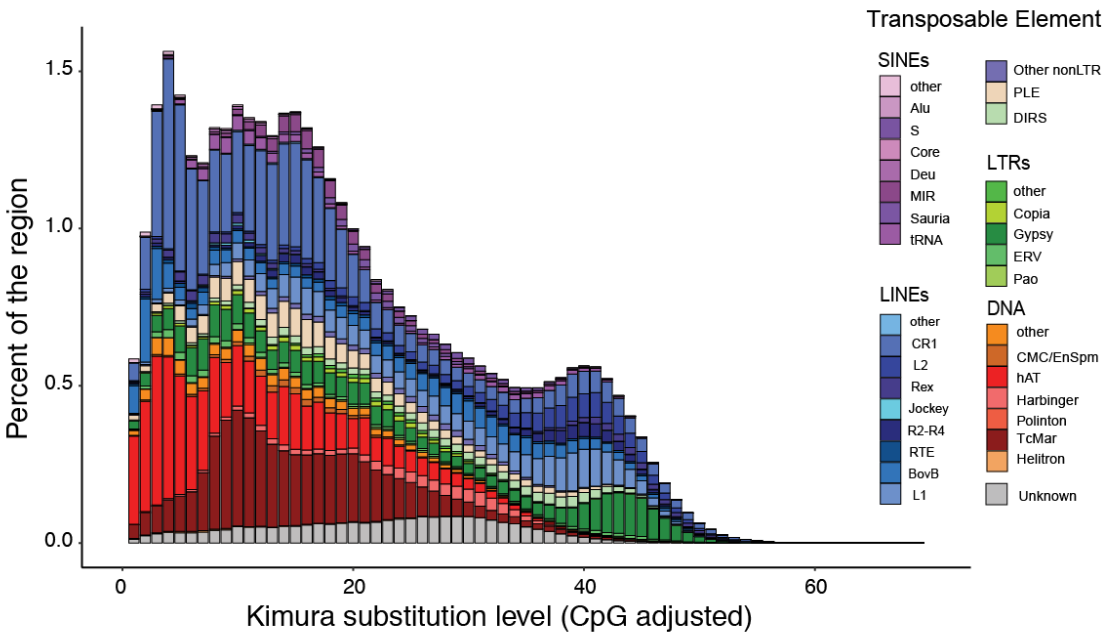
SUPPLEMENTARY FIGURES



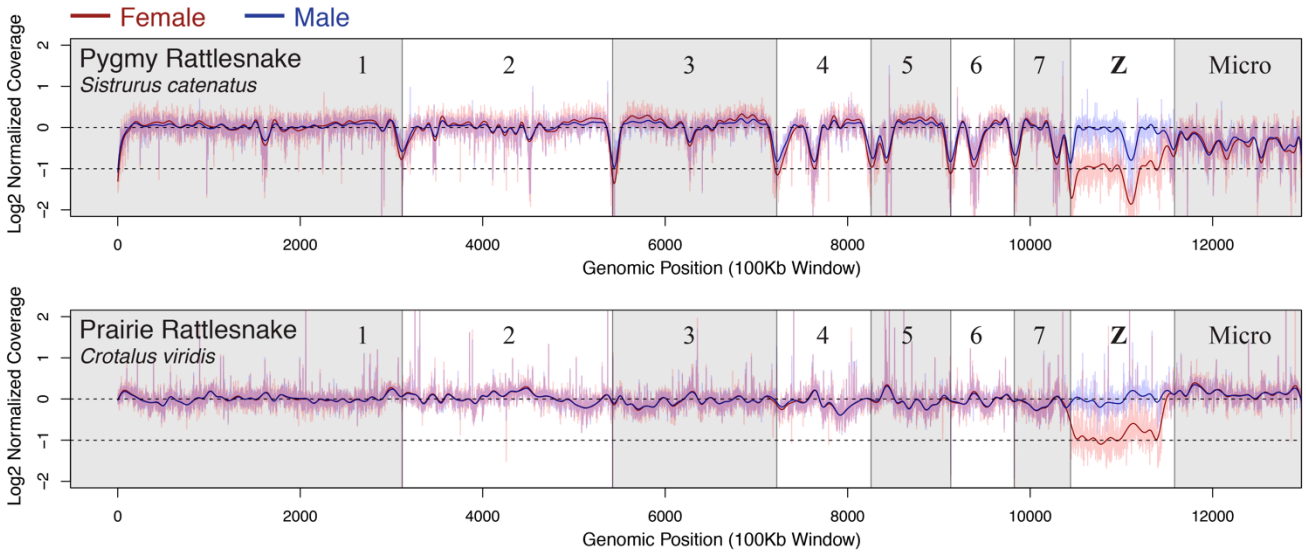
Supplementary Figure 1. Centromeric tandem repeat motif characterized using tandem repeats finder. Analysis of high frequency tandem repeats identified a 164-mer with high relative GC to the genomic background. The y-axis, tandem repeat mass, represents the relative abundance of tandem repeats of a given unit length and GC content.



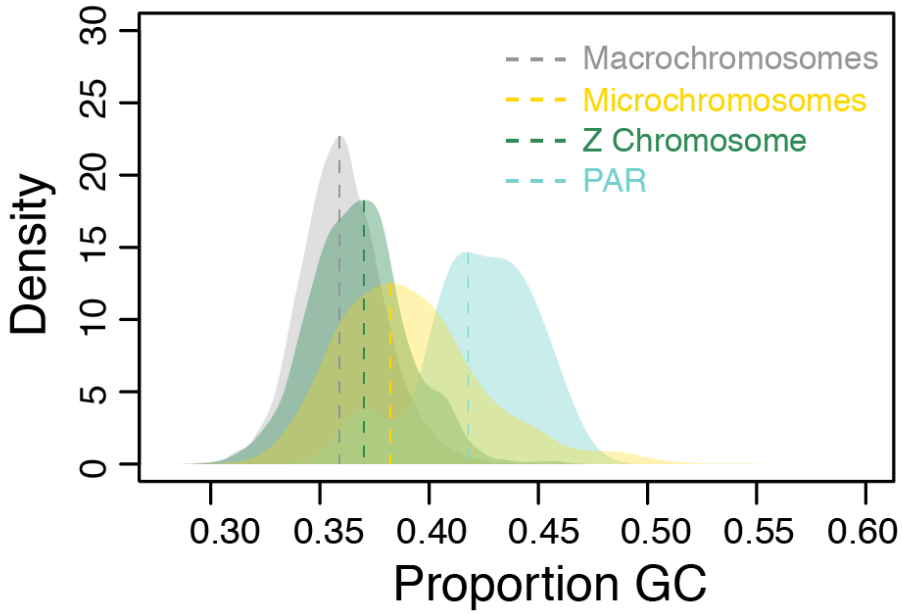
Supplementary Figure 2. Evolutionary patterns of genomic features of microchromosomes among reptiles. Values at nodes on the phylogenetic tree represent the node age in millions of years, and were obtained using median estimates from TimeTree. The heatmap to the right represents the relative abundance of a given measure on microchromosomes versus macrochromosomes within each species (blue values represent greater abundance on macrochromosomes and red values represent greater abundance on microchromosomes). Values in each heatmap cell equal the ratio of each measure on microchromosomes:macrochromosomes, and values with asterisks represent significant differences between microchromosomes and macrochromosomes.



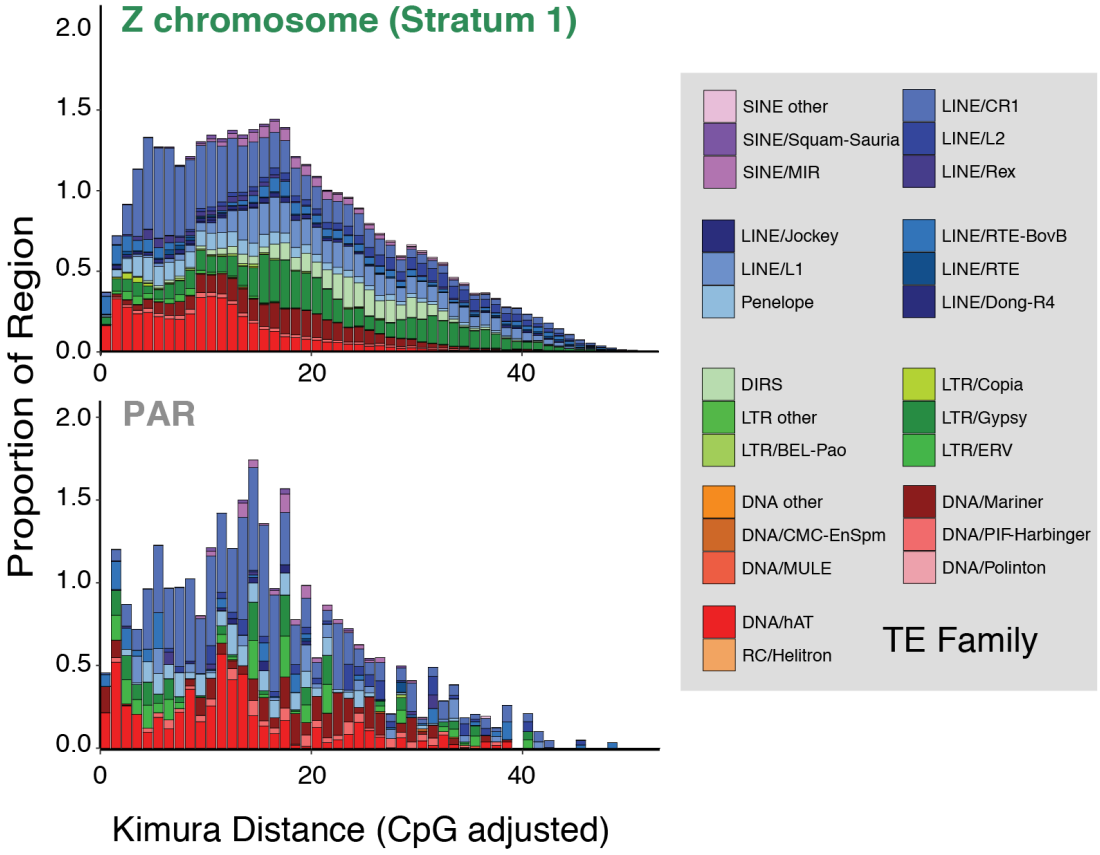
Supplementary Figure 3. Genomic repeat element abundance at a range of relative age values. Age is measured using the Kimura substitution level of transposable elements when compared to a consensus sequence.



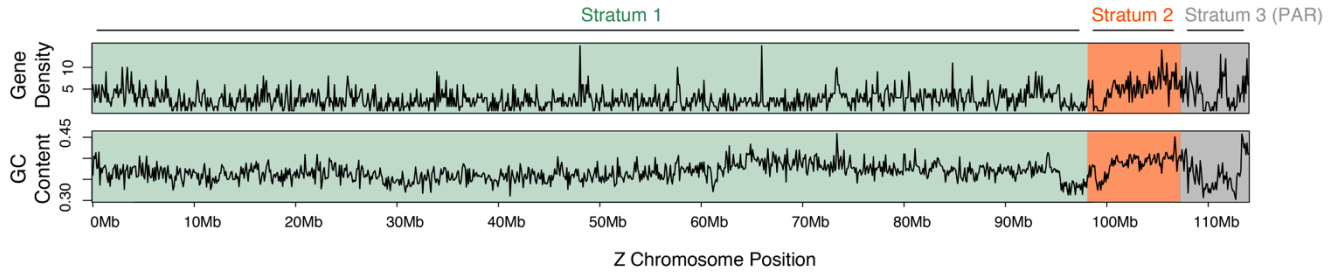
Supplementary Figure 4. Log2 normalized female (red) and male (blue) coverage of two rattlesnake species (*Sistrurus catenatus* and *Crotalus viridis*), when mapped to the prairie rattlesnake reference genome. The dashed line at zero represents the normalized coverage expectation for diploid loci, and the dashed line at -1 represents the expectation of a hemizygous locus. The transparent lines show values for each 100 Kb window in a sliding window analysis of coverage, and bold lines show a smoothened spline of relative coverage across the genome.



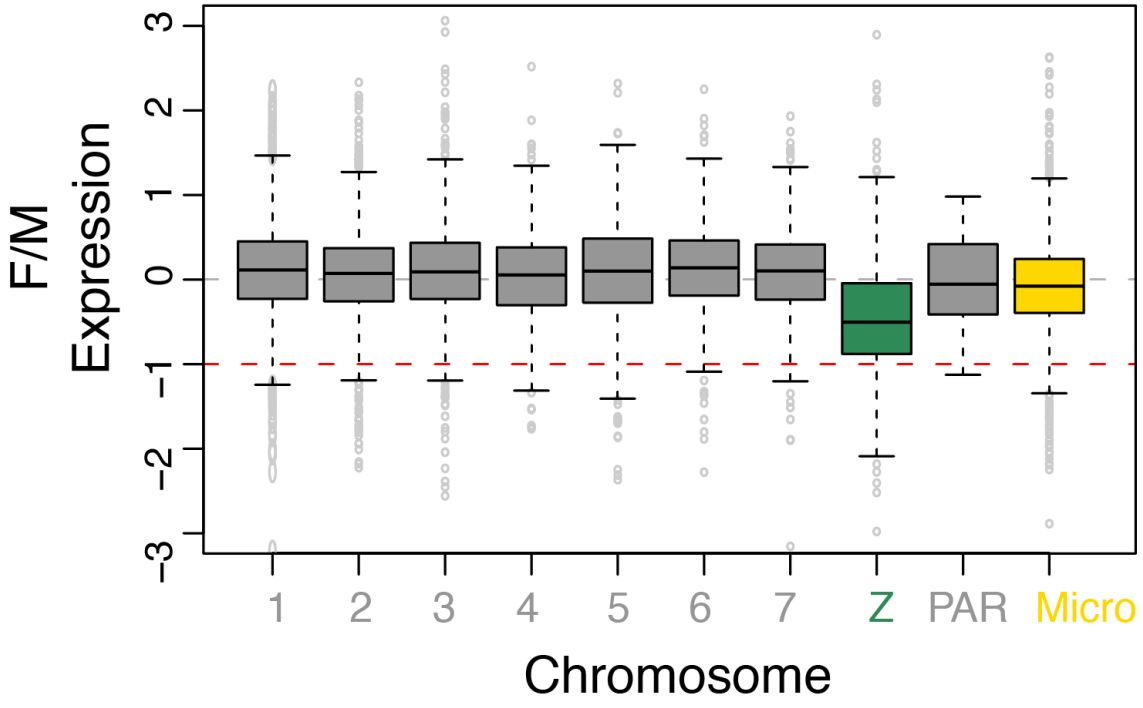
Supplementary Figure 5. Density distributions of GC content across prairie rattlesnake chromosomes, showing specific distributions of macrochromosomes, microchromosomes, the Z chromosome, and the pseudoautosomal region (PAR) of the sex chromosomes, specifically.



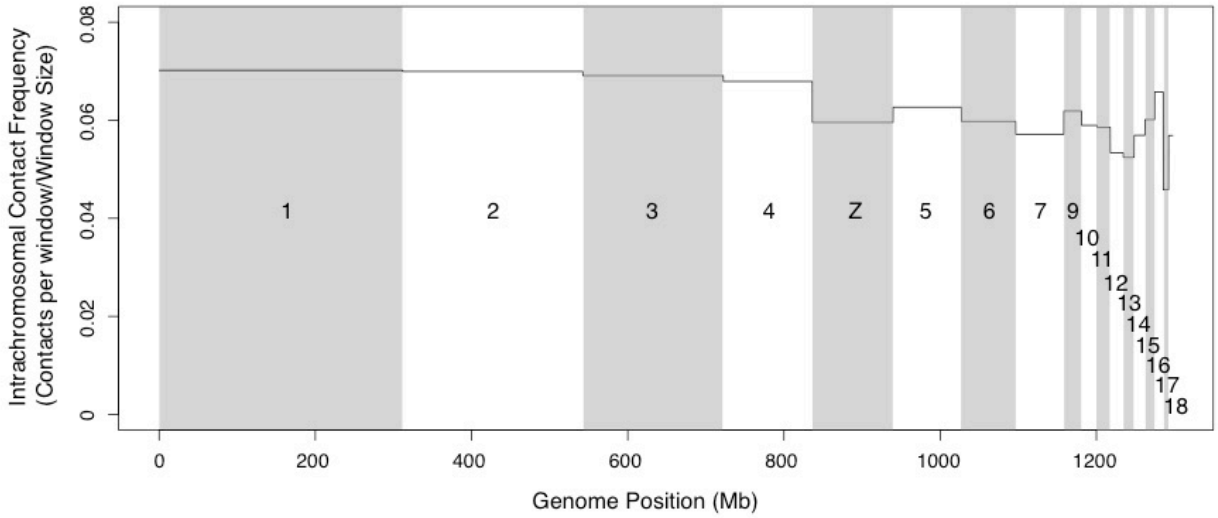
Supplementary Figure 6. Comparative age distributions of proportions of transposable elements (TEs) across Stratum 1 (upper) and the pseudoautosomal region (PAR; lower) of the rattlesnake Z chromosome.



Supplementary Figure 7. 100 Kb windowed scans of gene density (measured as number of genes per window) and GC content (i.e., proportion of GC bases within each window) across the Z chromosome of the prairie rattlesnake. The evolutionary strata are denoted by green (Stratum 1), orange (Stratum 2), and grey (Stratum 3; PAR) backgrounds.



Supplementary Figure 8. Boxplots of relative female:male gene expression on autosomal macrochromosomes (grey), the Z chromosome (green), the pseudoautosomal region (PAR; also grey), and microchromosomes. Outliers per chromosome are shown as small grey circles. The grey horizontal dashed line represents the expected value for autosomal loci, and the red dashed line represents gene expression for a hemizygous locus in the absence of any dosage compensation mechanisms.



Supplementary Figure 9. Intrachromosomal contact frequencies (i.e., the number of observed contacts between a chromosomal window and all other intrachromosomal windows divided by the size of the window) measured using Hi-C of the venom gland across rattlesnake chromosomes, demonstrating a constant level of intrachromosomal contact within each assembled chromosome. Individual chromosomes are labeled.

Supplementary Tables

Supplementary Table 1. Sequencing libraries used in the prairie rattlesnake genome assembly. Where noted, various libraries were used in the previous assembly (CroVir2.0), published in Pasquesi et al. (in review).

Library	Read Type	Number of Reads	Assembly Version
50bp short read	single end	9,536,384	CroVir2.0
100bp short read	paired end	449775645	CroVir2.0, CroVir3.0
150bp short read	paired end	41,211,014	CroVir2.0
150bp long insert mate pair (3-5Kb)	paired end	188,532,564	CroVir2.0
150bp long insert mate pair (6-8Kb)	paired end	189,928,342	CroVir2.0
PacBio long reads	-	1,027,365	CroVir2.0
Chicago long range proximity ligation library 1 (150bp)	paired end	251,689,106	CroVir3.0
Chicago long range proximity ligation library 2 (150bp)	paired end	206,176,028	CroVir3.0
Hi-C library 1 (150bp)	paired end	230,083,402	CroVir3.0
Hi-C library 2 (150bp)	paired end	160,673,944	CroVir3.0

Supplementary Table 2. Basic information about assembly versions for the prairie rattlesnake genome.

	Input Assembly (CroVir2.0)	Chicago Assembly	HiRise (Chicago + Hi-C) Assembly
Longest Scaffold (bp)	1,184,546	11,576,738	311,712,589
Number of Scaffolds	47,782	8,183	7,034
Number of Scaffolds > 1Kb	47,658	8,059	6,910
Contig N50 (Kb)	15.81	14.91	14.96
Scaffold N50 (Kb)	139	2,472	179,898
Number of Gaps	112,369	158,269	159,024
Percent of Genome in Gaps	5.84%	6.15%	6.16%

Supplementary Table 3. Genome-wide annotated repeat proportions identified using RepeatMasker.

	# elements	length masked (bp)	% of sequence	% element masked
Total masked	2966274	489373735	38.91	100.00
Total interspersed repeats	2348232	463237605	36.83	79.16
Retroelements	1139213	295244109	22.81	38.41
SINEs	173332	22894322	1.82	5.84
Squam1/Sauria	19230	3376458	0.27	0.65
Other SINEs	126898	15602678	1.24	4.28
LINEs	621859	170275973	13.54	20.96
CR1-Like	359387	91177000	7.25	12.12
CR1/L3	288888	74285822	5.91	9.74
L2	53219	12036490	0.96	1.79
Rex	19032	5339363	0.42	0.64
R1/LOA/Jockey	3272	854611	0.07	0.11
R2/R4/NeSL	35256	9045775	0.72	1.19
RTE/Bov-B	101958	32795496	2.61	3.44
L1/CIN4	78926	28358227	2.25	2.66
Other LINEs	154019	16472232	0.64	5.19
Other nonLTR	10119	1572442	0.13	0.34
DIRS	28657	13553057	1.08	0.97
PLEs	120162	19278497	1.53	4.05
LTR elements	156427	54116761	4.30	5.27
BEL/Pao	4007	1927682	0.15	0.14
Ty1/Copia	9160	3340874	0.27	0.31
Gypsy	77793	35080772	2.79	2.62
Retroviral	16727	5393228	0.43	0.56
Other LTR	48740	8374205	0.67	1.64
DNA transposons	850487	125287793	9.96	28.67
hobo-Activator	428247	60243144	4.79	14.44
Tc1-IS630-Pogo	283367	48888185	3.89	9.55
En-Spm	12485	1964905	0.16	0.42
MuDR-IS905	1300	383077	0.03	0.04
PiggyBac	131	22504	0.00	0.00
Tourist/Harbinger	80904	7193605	0.57	2.73
P elements	155	45074	0.00	0.01
Rolling-circles	3736	635885	0.05	0.13
SPIN	253	26640	0.00	0.01
Other DNA	39909	5884774	0.47	1.35
Unclassified	358532	48493199	3.86	12.09
Total interspersed repeats	2348232	463237605	36.83	79.16
Small RNA	2054	174940	0.01	0.07
Satellites	4952	1104344	0.09	0.17
Simple repeats	540288	28572170	2.27	18.21
Low complexity	70748	4755565	0.38	2.39

Supplementary Table 4. Mapping of cDNA markers from Matsubara et al. 2006 to the prairie rattlesnake genome. Locations of best BLAST hits of each cDNA marker to the genome are reported. Where noted, cDNA markers mapped with exceptional similarity to multiple locations in the genome, or did not map to the chromosome as predicted by Matsubara et al. 2006. Markers for which there were two high-similarity hits on multiple chromosomes are denoted with italics.

Marker	Accession	Chromosome	Scaffold	e-value	bit-score	Start Position	End Position	Notes
OMG	BW999947	1p	scaffold-ma1	6.00E-115	398	309337082	309336564	
XAB1	AU312353	1p	scaffold-ma1	2.00E-46	122	297437298	297437486	
MGC15407	AU312344	1p	scaffold-ma1	2.00E-65	92.3	288097081	288097206	
XPO1	AU312325	1p	scaffold-ma1	2.00E-113	153	289547707	289547901	
DEGS	AU312341	1p	scaffold-ma1	5.00E-106	356	269312409	269311948	
KIAA0007	AU312332	1p	scaffold-ma1	5.00E-50	120	265943692	265943841	
EPRS	AU312324	1p	scaffold-ma1	2.00E-91	174	270708945	270709160	
ARID4B	AU312346	1p	scaffold-ma1	1.00E-129	333	252059286	252059699	
QKI	AU312356	1p	scaffold-ma1	5.00E-112	124	246094729	246094887	
MDN1	AU312339	1p	scaffold-ma1	7.00E-60	109	211517498	211517349	
AFTIPHILIN	AU312311	1p	scaffold-ma1	5.00E-75	112	170752748	170752888	
SF3B1	AU312337	1q	scaffold-ma1	7.00E-95	215	150078848	150078576	
CACNB4	BW999948	1q	scaffold-ma1	1.00E-47	102	127283965	127283819	
ZFHX1B	BW999949	1q	scaffold-ma1	6.00E-93	204	123301385	123301101	
UMPS	AU312331	1q	scaffold-ma1	8.00E-95	198	113761458	113761724	
TCIRG1	BW999950	1q	scaffold-ma1	2.00E-72	164	102088882	102089094	
TSG101	AU312316	1q	scaffold-ma1	4.00E-76	113	88358887	88359054	
M11S1	AU312350	1q	scaffold-ma1	4.00E-31	94.5	70777673	70777560	
GPHN	AU312327	1q	scaffold-ma1	5.00E-68	116	60249829	60249644	
DNCH1	AU312310	1q	scaffold-ma1	1.00E-71	145	25060055	25059885	
HSPCA	BW999951	1q	scaffold-ma1	2.00E-123	149	25029984	25030184	
ISYNA1	AU312338	1q	scaffold-ma1	2.00E-89	178	7770987	7771196	
TUBGCP2	AU312343	1q	scaffold-ma1	4.00E-74	136	9697568	9697377	
ZFR	AU312309	2p	scaffold-ma2	8.00E-110	208	222653709	222653461	

PHAX	AU312322	2p	scaffold-ma2	3.00E-99	224	189308026	189307715	
VPS13A	BW999952	2p	scaffold-ma2	9.00E-70	109	179725513	179725656	
UBQLN1	BW999953	2p	scaffold-ma2	2.00E-87	132	182156077	182156238	
C9orf72	AU312326	2p	scaffold-ma2	5.00E-91	203	164760033	164760347	
KIAA0368	BW999954	2p	scaffold-ma2	1.00E-56	116	161287251	161287397	
TOPORS	BW999955	2p	scaffold-ma2	8.00E-118	410	162258381	162257809	
FAM48A	BW999956	2cen	scaffold-ma2	1.00E-45	102	157286823	157286680	
UNQ501	AU312305	2cen	scaffold-ma2	6.00E-118	284	142895238	142895636	
DCTN2	AU312317	2q	scaffold-ma2	4.00E-80	122	122527271	122527110	
EXOC7	BW999957	2q	scaffold-ma2	3.00E-93	121	92952368	92952526	
DDX5	BW999958	2q	scaffold-ma2	7.00E-112	144	108253948	108253775	
CCNG1	AU312308	2q	scaffold-ma2	6.00E-70	173	80553964	80553731	
CPEB4	AU312333	2q	scaffold-ma2	3.00E-119	250	72297563	72297874	
FLJ22318	AU312329	2q	scaffold-ma2	2.00E-105	194	51908839	51908582	
DCTN4	AU312349	2q	scaffold-ma2	4.00E-50	99.6	58962806	58962928	
C5orf14	AU312304	2q	scaffold-ma2	4.00E-120	329	64853582	64853127	
NOSIP	AU312303	2q	scaffold-Z	1.00E-51	93.6	92988551	92988661	Did not map to predicted chromosome
RBM5	BW999960	2q	scaffold-mi8	6.00E-78	90.4	9620291	9620181	Mapped to multiple chromosomes with high similarity
RBM5	BW999960	2q	scaffold-ma2	7.00E-13	76.1	130725514	130725606	Mapped to multiple chromosomes with high similarity
ITPR1	BW999961	2q	scaffold-ma2	9.00E-53	135	23858424	23858585	
ENPP2	BW999962	3p	scaffold-ma3	6.00E-90	121	9756367	9756209	
YWHAZ	BW999963	3p	scaffold-ma3	2.00E-99	180	16759896	16760114	
LRRCC1	BW999964	3p	scaffold-ma3	4.00E-83	150	21993774	21993565	
LYPLA1	BW999965	3p	scaffold-ma3	3.00E-107	149	31673258	31673440	
SS18	AU312302	3p	scaffold-ma3	1.00E-83	126	36811554	36811724	
MBP	AU312318	3p	scaffold-ma3	7.00E-111	179	49049170	49049382	
EPB41L3	BW999966	3p	scaffold-ma3	3.00E-84	141	40222999	40222808	
TUBB2A	BW999967	3p	scaffold-ma3	8.00E-91	155	59187732	59187532	
LRRC16	BW999968	3p	scaffold-ma3	2.00E-100	144	51025171	51025350	

SERPINB6	BW999969	3p	scaffold-ma5	5.00E-99	130	36540937	36540755	Mapped to multiple chromosomes with high similarity
SERPINB6	BW999969	3p	scaffold-ma3	2.00E-76	113	60484038	60483865	Mapped to multiple chromosomes with high similarity
BPHL	BW999970	3p	scaffold-ma3	1.00E-87	118	59199779	59199621	
KIF13A	BW999971	3p	scaffold-ma3	3.00E-78	139	53681516	53681349	
TPR	BW999972	3q	scaffold-ma3	6.00E-83	122	93408800	93408636	
AKR1A1	BW999973	3q	scaffold-ma3	9.00E-75	153	133869419	133869619	
ZNF326	BW999974	3q	scaffold-ma2	2.00E-77	120	224940437	224940586	Did not map to predicted chromosome
YIPF1	BW999975	3q	scaffold-ma3	6.00E-52	112	127724189	127724353	
BCAS2	AU312354	3q	scaffold-ma3	3.00E-51	141	151621402	151621229	
KIAA1219	BW999976	3q	scaffold-ma3	4.00E-101	158	155122635	155122844	
STAU1	BW999977	3q	scaffold-ma3	2.00E-116	169	165663812	165663594	
RBM12	BW999978	3q	scaffold-ma3	2.00E-152	406	154706304	154705780	
TPT1	BW999979	4p	scaffold-ma4	2.00E-68	148	1006155	1006349	
EIF2S3	AU312306	4p	scaffold-ma4	1.00E-111	126	49115724	49115885	
SYAP1	AU312328	4p	scaffold-ma4	3.00E-96	121	46147275	46147135	
DSCR3	AU312319	4q	scaffold-ma4	1.00E-74	119	60873037	60872873	
DCAMKL1	BW999980	4q	scaffold-ma4	8.00E-49	110	86291138	86291302	
ELMOD1	BW999981	4q	scaffold-ma4	1.00E-56	147	93207704	93207522	
BCCIP	AU312307	5q	scaffold-ma5	1.00E-46	148	32597249	32597061	
SH3MD1	AU312347	5q	scaffold-ma5	2.00E-119	378	45831798	45832379	
PPP1R7	BW999982	5q	scaffold-ma5	2.00E-92	228	56956062	56955736	
PDCD10	AU312342	5q	scaffold-ma5	4.00E-61	143	74805371	74805547	
TLOC1	AU312335	5q	scaffold-ma5	2.00E-45	101	76109988	76110125	
UCHL1	BW999983	6p	scaffold-ma7	4.00E-89	210	33298090	33298407	Did not map to predicted chromosome
GNAI2	BW999984	6p	scaffold-ma2	2.00E-106	126	49893686	49893841	Did not map to predicted chromosome
P4HB	BW999985	6p	scaffold-ma2	2.00E-69	100	97717890	97718012	Did not map to predicted chromosome
FLJ12571	AU312352	6q	scaffold-ma6	2.00E-46	117	46698606	46698752	
RANGAP1	AU312313	6q	scaffold-ma6	7.00E-71	95	47795604	47795500	
LDHB	BW999986	6q	scaffold-ma6	2.00E-60	117	69268248	69268418	

SEC3L1	AU312345	7p	scaffold-ma7	3.00E-58	125	55644074	55643916	
KIAA1109	AU312348	7q	scaffold-ma7	2.00E-60	124	30398905	30398711	
RAP1GDS1	AU312351	7q	scaffold-ma7	2.00E-91	112	12141068	12140931	
GAD2	BW999991	Zp	scaffold-Z	1.00E-109	136	17484512	17484336	
WAC	AU312355	Zp	scaffold-Z	3.00E-93	209	16303681	16303947	
KLF6	BW999992	Zp	scaffold-ma2	1.00E-99	366	47130305	47130796	Did not map to predicted chromosome
LOC90693	BW999993	Zp	scaffold-ma7	4.00E-127	301	34444161	34444577	Mapped to multiple chromosomes with high similarity
LOC90693	BW999993	Zp	scaffold-Z	1.00E-107	291	34827559	34827182	Mapped to multiple chromosomes with high similarity
TAX1BP1	AU312320	Zp	scaffold-Z	1.00E-86	141	36989995	36990174	
RAB5A	BW999994	Zp	scaffold-Z	9.00E-94	166	40227424	40227215	
CTNNB1	BW999995	Zcen	scaffold-Z	3.00E-129	275	49548885	49549226	
AMPH	BW999996	Zcen	scaffold-Z	1.00E-66	101	55612836	55612955	
TUBG1	BW999997	Zq	scaffold-Z	5.00E-89	116	17359265	17359113	
GH1	BW999998	Zq	scaffold-Z	2.00E-115	179	77397011	77396727	
MYST2	BW999999	Zq	scaffold-Z	6.00E-122	293	90785118	90784714	
NEF3	BW999987	micro	scaffold-mi1	1.00E-102	352	13833430	13832942	
ASB6	AU312340	micro	scaffold-mi7	1.00E-95	161	6270589	6270353	
RPL12	BW999988	micro	scaffold-mi7	6.00E-67	95.5	7974658	7974542	
FLJ25530	AU312336	micro	scaffold-mi1	4.00E-98	255	8157147	8156806	
HSPA8	BW999989	micro	scaffold-ma1	2.00E-124	236	20422342	20422662	Mapped to multiple chromosomes with high similarity
HSPA8	BW999989	micro	scaffold-mi1	3.00E-123	259	2089357	2089025	Mapped to multiple chromosomes with high similarity
GLCE	AU312330	micro	scaffold-mi10	1.00E-79	234	24861	24577	
POLG	AU312315	micro	scaffold-mi3	4.00E-97	116	10042696	10042845	
LOC283820	AU312323	micro	scaffold-mi5	8.00E-71	116	3659851	3659708	
PARN	AU312312	micro	scaffold-mi7	1.00E-66	73.9	12029447	12029361	
ATRX	BW999990	micro	scaffold-mi4	3.00E-63	102	1268001	1268126	

Supplementary Table 5. Transcription factors significantly upregulated in the venom gland.

Gene ID	Rattlesnake Gene Detail
ATF6	augustus_masked-scaffold-ma3-processed-gene-300.3
ELF5	maker-scaffold-ma1-augustus-gene-235.5
FOXC2	augustus_masked-scaffold-mi6-processed-gene-2.1
CREB3L2	maker-scaffold-ma6-augustus-gene-195.2
HSP90B1	maker-scaffold-ma6-augustus-gene-185.14
GRHL1	maker-scaffold-ma1-augustus-gene-601.8
NCOA2	maker-scaffold-ma3-augustus-gene-89.6
NFIA	maker-scaffold-ma3-augustus-gene-414.2
NFIB	maker-scaffold-ma2-augustus-gene-569.3
NFIB	maker-scaffold-ma2-augustus-gene-569.2
NFIX	maker-scaffold-ma2-augustus-gene-473.3
NR4A2	maker-scaffold-ma1-augustus-gene-428.4
SREBF2	maker-scaffold-ma6-augustus-gene-158.15

Supplementary Table 6. RNAseq libraries used in this study.

Sample ID	Tissue	Raw Reads	Quality Trimmed Reads
CroVirPan	pancreas	28,126,703	27,073,946
CroVirTon	tongue	24,451,116	23,561,349
CroVirVG1	venom gland	41,744,110	40,147,306
CroVirVG3	venom gland	29,216,664	28,035,353
Cvv01	liver	7,833,506	7,365,740
Cvv02	liver	7,451,792	7,064,234
Cvv11	liver	9,218,939	8,441,587
Cvv20	kidney	6,958,120	6,580,387
Cvv22	kidney	8,116,679	7,601,517
Cvv23	kidney	7,193,762	6,785,947
Cvv25	skin	7,849,895	7,303,441
Cvv26	pancreas	8,886,612	8,160,214
Cvv27	venom gland	3,098,151	2,928,974
Cvv28	lung	6,613,196	6,024,613
Cvv29	testes	5,055,189	4,745,375
Cvv30	accessory venom gland	3,261,326	3,053,142
Cvv31	shaker muscle	4,290,989	3,996,274
Cvv32	pancreas	4,836,715	4,566,165
Cvv33	brain	3,815,570	3,569,113
Cvv34	stomach	5,297,110	4,993,142
Cvv35	ovaries	3,737,870	3,528,104
Cvv36	rictal gland	6,654,626	6,070,883
Cvv37	spleen	7,776,020	6,975,210
Cvv38	blood	2,550,433	2,364,162

Supplementary Table 7. Details of Illumina Nextera resequencing libraries used for comparative female/male read coverage across the rattlesnake genome.

Library Type	Read Length	Sample ID	Species	Sex	Number of Mapped Reads
Illumina Nextera	150 bp paired end	CV0007	*Crotalus viridis viridis*	Male	20,279,801
Illumina Nextera	150 bp paired end	CV0011	*Crotalus viridis viridis*	Female	4,975,491

Supplementary Table 8. GC variation in windows of various sizes for 12 squamate species. Values for each species are measured as the standard deviation (SD) of GC content in all sampled windows of a given size. Information for 5, 20, and 80 Kb windows are also presented in Fig. 1c. Missing data (i.e., window sizes that were too large and contained greater than the threshold allowed missing data) are denoted with '-'.

Window Size (bp)	*Gekko japonicus*	*Eublepharis macularius*	*Ophisaurus gracilis*	*Shinisaurus crocodilurus*	*Pogona vitticeps*	*Anolis carolinensis*
5,000	0.039295606	0.037140406	0.037038224	0.03488877	0.03681681	0.032312269
20,000	0.028980944	0.027338004	0.029217483	0.027425317	0.030930264	0.021209
40,000	0.025219459	0.024838347	0.027141528	0.025322106	0.029367252	0.017608402
80,000	0.021385708	0.023326607	0.025558162	0.023843432	0.028238318	0.015121097
160,000	0.01811246	0.022646783	0.024536212	0.022632678	0.027330318	0.013089382
240,000	-	0.022203903	0.023356372	0.021943776	0.026943855	0.012088733
320,000	-	0.022121291	0.022899173	0.021312719	0.026617904	0.011287772

Window Size (bp)	*Boa constrictor*	*Python molurus*	*Ophiophagus hannah*	*Thamnophis sirtalis*	*Deinagkistrodon acutus*	*Crotalus viridis*
5,000	0.043942864	0.042024505	0.040098669	0.047076022	0.047062019	0.041210929
20,000	0.034934365	0.035837726	0.031894398	0.037865804	0.03882085	0.032232558
40,000	0.030576918	0.033337717	0.028952912	0.03429097	0.036517713	0.029884634
80,000	0.023292703	0.030197592	0.026685436	0.031202717	0.034964163	0.0281043
160,000	0.014736549	0.02736241	0.024597185	0.02894796	0.033486765	0.026806291
240,000	-	0.024725646	0.023968494	0.026250057	0.032562166	0.02616041
320,000	-	0.023707617	0.023468328	0.024606171	0.031784231	0.025840409

Supplementary Table 9. Representative sequences for known snake venom gene families used to annotate venom genes in the rattlesnake genome.

Gene Family	Accession	Sequence Type	Species
5'Nucleotidase	AK291667.1	mRNA	*Homo sapiens*
Acetylcholinesterase	U54591.1	mRNA	*Bungarus fasciatus*
AVItoxin	EU195459.1	mRNA	*Varanus komodoensis*
C-type Lectin	JF895761.1	mRNA	*Crotalus oreganus helleri*
Cobra Venom Factor	U09969.2	mRNA	*Naja kaouthia*
CRISp (cysteine-rich secretory protein)	HQ414088.1	mRNA	*Crotalus adamanteus*
Cystatin	FJ411289.1	mRNA	*Naja kaouthia*
Extendin	EU790960.1	mRNA	*Heloderma suspectum*
Exonuclease	XM_015826835.1	mRNA	*Protobothrops mucrosquamatus*
Hyaluronidase	HQ414098.1	mRNA	*Crotalus adamanteus*
LAAO (L-amino acid oxidase)	HQ414099.1	mRNA	*Crotalus adamanteus*
SVMP I (class I snake venom metalloproteinase)	HM443635.1	mRNA	*Bothrops neuwiedi*
SVMP II (class II snake venom metalloproteinase)	HM443637.1	mRNA	*Bothrops neuwiedi*
SVMP III (class III snake venom metalloproteinase)	HM443632.1	mRNA	*Bothrops neuwiedi*
Nerve growth factor	AF306533.1	mRNA	*Crotalus durissus terrificus*
Phosphodiesterase	HQ414102.1	mRNA	*Crotalus adamanteus*
PLA2_I (vipers)	AF403134.1	mRNA	*Crotalus viridis viridis*
PLA2_II (elapids)	GU190815.1	mRNA	*Bungarus flaviceps*
Sarafotoxin	L07528.1	mRNA	*Atractaspis engaddensis*
Serine Proteinase	HQ414121.1	mRNA	*Crotalus adamanteus*
3FTX (Three-finger Toxin)	DQ273582.1	mRNA	*Ophiophagus hannah*
Veficolin	GU065323.1	mRNA	*Cerberus rynchops*
VEGF (Vascular Endothelial Growth Factor)	AB848141.1	mRNA	*Protobothrops mucrosquamatus*
Vespryn	EU401840.1	mRNA	*Oxyuranus scutellatus*
Waprin	EU401843.1	mRNA	*Oxyuranus scutellatus*
Kunitz (serine peptidase inhibitor, Kunitz type)	JU173666.1	mRNA	*Crotalus adamanteus*

Thrombin-like (thrombin-like venom gland enzyme)	AJ001209.1	mRNA	*Deinagkistrodon acutus*
Ficolin	GBUG01000048.1	mRNA	*Echis coloratus*
Disintegrin	AJ131345.1	mRNA	*Deinagkistrodon acutus*
FactorV (venom coagulation factor V)	XM_015815922.1	mRNA	*Protobothrops mucrosquamatus*
FactorX	XM_015819885.1	mRNA	*Protobothrops mucrosquamatus*
Prokineticin	XM_015822870.1	mRNA	*Protobothrops mucrosquamatus*
Ohanin (ohanin-like)	XM_015818414.1	mRNA	*Protobothrops mucrosquamatus*
Complement C3 (Cadam VF)	JU173742.1	mRNA	*Crotalus adamanteus*
Crotasin	AF250212.1	mRNA	*Crotalus durissus terrificus*
Endothelin	XM_015810852.1	mRNA	*Protobothrops mucrosquamatus*
Kallikrein	GALC01000005.1	mRNA	*Crotalus oreganus helleri*
Lynx1 (Ly6/neurotoxin 1)	XM_014066791.1	mRNA	*Thamnophis sirtalis*
Natriuretic Peptide (bradykinin potentiating peptide and C-type natriuretic peptide precursor isoform 2)	AF308594.2	mRNA	*Crotalus durissus terrificus*
sPla/ryanodine receptor	XM_015823102.1	mRNA	*Protobothrops mucrosquamatus*
WAP four-disulfide core domain protein 5 (Whey Acidic Protein/secretory leuki proteinase inhibitor)	XM_015822353.1	mRNA	*Protobothrops mucrosquamatus*
Myotoxin	HQ414100.1	mRNA	*Crotalus adamanteus*
PLA2	APD70899.1	protein	*Crotalus atrox*
SVMP	Q90282.1	protein	*Crotalus atrox*
Serine Proteinase	F8S114.1	protein	*Crotalus adamanteus*

Supplementary Table 10. Annotated venom gene homologs in the prairie rattlesnake genome. Genes were annotated using materials detailed in Supplementary Table 9.

Venom Gene Family	Rattlesnake Scaffold	Start Position (bp)	End Position (bp)
3-Finger toxin	scaffold-ma1	103004868	103021927
3-Finger toxin	scaffold-ma1	102999393	103000958
5' Nucleotidase	scaffold-ma5	46133017	46179118
5' Nucleotidase	scaffold-ma6	55711914	55732365
5' Nucleotidase	scaffold-mi1	18004217	18021456
5' Nucleotidase	scaffold-ma2	45090212	45121335
5' Nucleotidase	scaffold-ma2	134237148	134264183
Acetylcholinesterase	scaffold-ma2	4047955	4053281
Acetylcholinesterase	scaffold-ma2	3948506	3952373
Acetylcholinesterase	scaffold-ma2	4016363	4018146
Acetylcholinesterase	scaffold-ma2	4026170	4045822
Acetylcholinesterase	scaffold-ma5	73971094	73976212
Acetylcholinesterase	scaffold-ma5	74015346	74036663
Acetylcholinesterase	scaffold-un210	16032	17552
Bradykinin potentiating and natriuretic peptide	scaffold-un187	22386	23524
C-type lectin	scaffold-mi5	3276042	3284747
C-type lectin	scaffold-mi5	11650747	11653723
C-type lectin	scaffold-Z	21883578	21895509
C-type lectin	scaffold-Z	21706900	21776775
C-type lectin	scaffold-Z	21786524	21797211
C-type lectin	scaffold-Z	108214710	108236532
Cysteine-rich secretory protein	scaffold-ma1	169434958	169437996
Cysteine-rich secretory protein	scaffold-ma1	169423774	169434684
Cysteine-rich secretory protein	scaffold-ma3	25391938	25416947
Cysteine-rich secretory protein	scaffold-mi6	1021447	1040191
Exonuclease	scaffold-mi7	8097114	8103411

Exonuclease	scaffold-ma1	5804894	5842638
Exonuclease	scaffold-mi3	10271502	10274220
Exonuclease	scaffold-ma6	12590208	12591465
Factor V	scaffold-mi4	8493826	8518402
Factor V	scaffold-mi4	8479637	8493564
Factor V	scaffold-ma4	81074882	81113119
Glutaminyl cyclase	scaffold-ma1	256551622	256564040
Glutaminyl cyclase	scaffold-mi7	5091107	5094268
Hyaluronidase	scaffold-ma6	14952252	14955850
Hyaluronidase	scaffold-ma2	45901201	45920587
Hyaluronidase	scaffold-ma2	49137409	49145188
Hyaluronidase	scaffold-ma2	49106981	49118469
Kunitz peptide	scaffold-mi7	3590975	3597607
Kunitz peptide	scaffold-mi8	4992795	5002390
L-amino acid oxidase	scaffold-ma4	56914906	56948498
L-amino acid oxidase	scaffold-ma4	85461961	85468906
L-amino acid oxidase	scaffold-ma2	4658599	4661642
L-amino acid oxidase	scaffold-ma2	4654769	4658293
Myotoxin/crotamine	scaffold-ma1	289328153	289328605
Nerve growth factor	scaffold-Z	93342025	93347811
Nerve growth factor	scaffold-ma1	76711308	76727703
PLA2	scaffold-mi7	3019970	3021876
PLA2	scaffold-mi7	3027607	3029199
PLA2	scaffold-mi7	3031464	3033348
PLA2	scaffold-mi7	3037103	3038488
PLA2	scaffold-mi7	3042118	3043697
Serine Proteinase	scaffold-mi2	8569773	8575182
Serine Proteinase	scaffold-mi2	8588278	8593660
Serine Proteinase	scaffold-mi2	8628274	8636651

Serine Proteinase	scaffold-mi2	8664603	8670797
Serine Proteinase	scaffold-mi2	8739986	8745649
Serine Proteinase	scaffold-mi2	8752578	8759324
Serine Proteinase	scaffold-mi2	8864675	8879153
Serine Proteinase	scaffold-mi2	8937526	8947481
Serine Proteinase	scaffold-mi2	8960028	8980478
Snake venom metalloproteinase	scaffold-mi1	13901629	14014239
Snake venom metalloproteinase	scaffold-mi1	14022082	14075370
Snake venom metalloproteinase	scaffold-mi1	14091987	14112667
Snake venom metalloproteinase	scaffold-mi1	14147865	14170405
Snake venom metalloproteinase	scaffold-mi1	14174872	14190142
Snake venom metalloproteinase	scaffold-mi1	14211673	14242249
Snake venom metalloproteinase	scaffold-mi1	14248933	14272689
Snake venom metalloproteinase	scaffold-mi1	14281564	14300774
Snake venom metalloproteinase	scaffold-mi1	14368422	14393313
Snake venom metalloproteinase	scaffold-mi1	14401627	14424637
Snake venom metalloproteinase	scaffold-mi1	14310844	14338336
Veficolin/Ficolin	scaffold-mi7	5271880	5282014
Veficolin/Ficolin	scaffold-ma3	179788950	179790745
Veficolin/Ficolin	scaffold-ma1	232337083	232340714
Veficolin/Ficolin	scaffold-ma1	232312034	232335439
Vascular endothelial growth factor	scaffold-ma7	40288572	40327884
Vascular endothelial growth factor	scaffold-ma1	40733075	40747358
Vascular endothelial growth factor	scaffold-ma1	260248287	260272500
Venom Factor	scaffold-Z	79798672	79803249
Venom Factor	scaffold-Z	79749464	79761456
Venom Factor	scaffold-ma2	1573588	1616446
Venom Factor	scaffold-ma2	137559964	137560374
Venom Factor	scaffold-ma2	137553669	137558461

Venom Factor	scaffold-ma2	137623562	137648584
Venom Factor	scaffold-ma2	137651285	137653877
Venom Factor	scaffold-ma2	137710627	137728987
Venom Factor	scaffold-ma2	137753804	137775039
Venom Factor	scaffold-ma2	137735629	137741352
Vespryn/Ohanin	scaffold-ma2	4377779	4385668
Vespryn/Ohanin	scaffold-ma2	109834300	109838076
Waprin	scaffold-ma1	204655764	204666466

REFERENCES

Adams, R. H., D. R. Schield, D. C. Card, H. Blackmon, and T. A. Castoe. 2016. GppFst: Genomic posterior predictive simulations of FST and dXY for identifying outlier loci from population genomic data. *Bioinformatics* 33:1414-1415.

Alfoldi, J., F. Di Palma, M. Grabherr, C. Williams, L. Kong, E. Mauceli, P. Russell, C. B. Lowe, R. E. Glor, J. D. Jaffe, et al. 2011. The genome of the green anole lizard and a comparative analysis with birds and mammals. *Nature* 477:587-591.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215:403-410.

Anders, S., D. J. McCarthy, Y. Chen, M. Okoniewski, G. K. Smyth, W. Huber, and M. D. Robinson. 2013. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nature Protocols* 8:1765.

Anderson, C. G. and E. Greenbaum. 2012. Phylogeography of northern populations of the black-tailed rattlesnake (*Crotalus Molossus* Baird and Girard, 1853), with the revalidation of *C. ornatus* Hallowell, 1854. *Herpetological Monographs* 26:19-57.

Andrew, A. L., B. W. Perry, D. C. Card, D. R. Schield, R. P. Ruggiero, S. E. McGaugh, A. Choudhary, S. M. Secor, and T. A. Castoe. 2017. Growth and stress response mechanisms underlying post-feeding regenerative organ growth in the Burmese python. *BMC Genomics* 18:338.

Arevalo, E., S. K. Davis, and J. W. Sites. 1994. Mitochondrial-DNA Sequence Divergence and Phylogenetic-Relationships among 8 Chromosome Races of the *Sceloporus grammicus* Complex (Phrynosomatidae) in Central Mexico. *Systematic Biology* 43:387-418.

Arnold, C. 2016. The snakebite fight. *Nature* 537:26-28.

Ashton, K. G. and A. de Queiroz. 2001. Molecular systematics of the western rattlesnake, *Crotalus viridis* (Viperidae), with comments on the utility of the D-loop in phylogenetic studies of snakes. *Molecular Phylogenetics and Evolution* 21:176-189.

Avise, J. C. 2000. Phylogeography: The History and Formation of Species. Harvard University Press, Cambridge, MA.

Avise, J. C. and R. C. Vrijenhoek. 1987. Mode of inheritance and variation of mitochondrial-DNA in hybridogenetic fishes of the genus *Poeciliopsis*. *Molecular Biology and Evolution* 4:514-525.

Baker, R. J., J. J. Bull, and G. A. Mengden. 1972. Karyotypic Studies of 38 Species of North-American Snakes. *Copeia* 1972:257-265.

Bandelt, H. J., P. Forster, and A. Rohl. 1999. Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution* 16:37-48.

Bao, W., K. K. Kojima, and O. Kohany. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* 6:11.

Beckstette, M., R. Homann, R. Giegerich, and S. Kurtz. 2006. Fast index based algorithms and software for matching position specific scoring matrices. *BMC Bioinformatics* 7:389.

Bellott, D. W., H. Skaletsky, T.-J. Cho, L. Brown, D. Locke, N. Chen, S. Galkina, T. Pyntikova, N. Koutseva, and T. Graves. 2017. Avian W and mammalian Y chromosomes convergently retained dosage-sensitive regulators. *Nature Genetics* 49:387-394.

Blair, C. and S. Sanchez-Ramirez. 2016. Diversity-dependent cladogenesis throughout western Mexico: Evolutionary biogeography of rattlesnakes (Viperidae: Crotalinae: *Crotalus* and *Sistrurus*). *Molecular Phylogenetics and Evolution* 97:145-154.

Blanchette, M., W. J. Kent, C. Riemer, L. Elnitski, A. F. A. Smit, K. M. Roskin, R. Baertsch, K. Rosenbloom, H. Clawson, and E. D. Green. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Research* 14:708-715.

Boetzer, M., C. V. Henkel, H. J. Jansen, D. Butler, and W. Pirovano. 2010. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 27:578-579.

Bogdanova, V. S. 2007. Inheritance of organelle DNA markers in a pea cross associated with nuclear-cytoplasmic incompatibility. *Theoretical and Applied Genetics* 114:333-339.

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

Bouckaert, R., J. Heled, D. Kuhnert, T. Vaughan, C. H. Wu, D. Xie, M. A. Suchard, A. Rambaut, and A. J. Drummond. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Computational Biology* 10:e1003537.

Boughman, J. W. 2001. Divergent sexual selection enhances reproductive isolation in sticklebacks. *Nature* 411:944-948.

Braconnot, P., B. Otto-Bliesner, S. Harrison, S. Joussaume, J. Y. Peterchmitt, A. Abe-Ouchi, M. Crucifix, E. Driesschaert, T. Fichefet, C. D. Hewitt, et al. 2007. Results of PMIP2 coupled simulations of the Mid-Holocene and Last Glacial Maximum - Part 1: experiments and large-scale features. *Past Climatic Variability* 3:261-277.

Bradnam, K. R., J. N. Fass, A. Alexandrov, P. Baranay, M. Bechner, I. Birol, S. Boisvert, J. A. Chapman, G. Chapuis, R. Chikhi, et al. 2013. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience* 2:10.

Bryant, D., R. Bouckaert, J. Felsenstein, N. A. Rosenberg, and A. RoyChoudhury. 2012. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Molecular Biology and Evolution* 29:1917-1932.

Bryson, R. W., U. O. García‐Vázquez, and B. R. Riddle. 2011. Phylogeography of Middle American gophersnakes: mixed responses to biogeographical barriers across the Mexican Transition Zone. *Journal of Biogeography* 38:1570-1584.

Burbrink, F. T. and T. A. Castoe. 2009. Molecular phylogeography of snakes *in* R. Seigel, and S. Mullin, eds. *Snakes: Ecology and Conservation*. Cornell University Press, Ithaca, NY.

Burnham, K. P. and D. R. Anderson. 2003. Model selection and multimodel inference: a practical information-theoretic approach. Springer Science & Business Media.

Burton, R. S., C. K. Ellison, and J. S. Harrison. 2006. The sorry state of F2 hybrids: consequences of rapid mitochondrial DNA evolution in allopatric populations. *The American Naturalist* 168:S14-S24.

Bushnell, B. 2014. BBMap: a fast, accurate, splice-aware aligner. Ernest Orlando Lawrence Berkeley National Laboratory, Berkeley, CA (US).

Campbell, J. A. and W. W. Lamar. 2004. The Venomous Reptiles of the Western Hemisphere. Cornell University Press, Ithaca, NY.

Cantarel, B. L., I. Korf, S. M. C. Robb, G. Parra, E. Ross, B. Moore, C. Holt, A. S. Alvarado, and M. Yandell. 2008. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research* 18:188-196.

Cardwell, M. D. 2008. The reproductive ecology of Mohave rattlesnakes. *Journal of Zoology* 274:65-76.
Caro, S. P., A. Charmantier, M. M. Lambrechts, J. Blondel, J. Balthazart, and T. D. Williams. 2009. Local adaptation of timing of reproduction: females are in the driver's seat. *Functional Ecology* 23:172-179.
Casewell, N. R., S. C. Wagstaff, R. A. Harrison, C. Renjifo, and W. Wüster. 2011. Domain loss facilitates accelerated evolution and neofunctionalization of duplicate snake venom metalloproteinase toxin genes. *Molecular Biology and Evolution* 28:2637-2649.
Caseys, C., K. N. Stölting, T. Barbará, S. C. González-Martínez, and C. Lexer. 2015. Patterns of genetic diversity and differentiation in resistance gene clusters of two hybridizing European *Populus* species. *TREE Genetics & Genomes* 11:81.
Castoe, T. A., A. P. J. de Koning, K. T. Hall, D. C. Card, D. R. Schield, M. K. Fujita, R. P. Ruggiero, J. F. Degner, J. M. Daza, W. J. Gu, et al. 2013. The Burmese python genome reveals the molecular basis for extreme adaptation in snakes. *Proceedings of the National Academy of Sciences of the United States of America* 110:20645-20650.
Castoe, T. A., K. T. Hall, M. L. Guibotsy Mboulas, W. Gu, A. P. de Koning, S. E. Fox, A. W. Poole, V. Vemulapalli, J. M. Daza, T. Mockler, et al. 2011. Discovery of highly divergent repeat landscapes in snake genomes using high-throughput sequencing. *Genome Biology and Evolution* 3:641-653.
Castoe, T. A., Z. J. Jiang, W. Gu, Z. O. Wang, and D. D. Pollock. 2008. Adaptive evolution and functional redesign of core metabolic proteins in snakes. *PloS One* 3:e2201.
Castoe, T. A. and C. L. Parkinson. 2006. Bayesian mixed models and the phylogeny of pitvipers (Viperidae: Serpentes). *Molecular Phylogenetics and Evolution* 39:91-110.
Castoe, T. A., C. L. Spencer, and C. L. Parkinson. 2007. Phylogeographic structure and historical demography of the western diamondback rattlesnake (*Crotalus atrox*): A perspective on North American desert biogeography. *Molecular Phylogenetics and Evolution* 42:193-212.
Catalano, D., F. Licciulli, A. Turi, G. Grillo, C. Saccone, and D. D'Elia. 2006. MitoRes: a resource of nuclear-encoded mitochondrial genes and their products in Metazoa. *BMC Bioinformatics* 7:36.
Catchen, J., P. A. Hohenlohe, S. Bassham, A. Amores, and W. A. Cresko. 2013. Stacks: an analysis tool set for population genomics. *Molecular Ecology* 22:3124-3140.
Cate, R. L. and A. L. Bieber. 1978. Purification and characterization of Mojave rattlesnake (*Crotalus scutulatus scutulatus*) toxin and its subunits. *Archives of Biochemistry and Biophysics* 189:397-408.
Chapman, J. A., I. Ho, S. Sunkara, S. Luo, G. P. Schroth, and D. S. Rokhsar. 2011. Meraculous: de novo genome assembly with short paired-end reads. *PloS One* 6:e23501.
Chikhirzhina, G. I., R. I. Al'-Shekhadat, and E. V. Chikhirzhina. 2008. Transcription factors of the nuclear factor 1 (*NF1*) family. Role in chromatin remodelation. *Molekuliarnaia biologiia* 42:388-404.
Cohn, M. J. and C. Tickle. 1999. Developmental basis of limblessness and axial patterning in snakes. *Nature* 399:474-479.
Consortium, I. H. G. S. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860-921.
Coyne, J. A. and H. A. Orr. 2004. Speciation. Sinauer Associates, Sunderland, MA.

Cruickshank, T. E. and M. W. Hahn. 2014. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology* 23:3133-3157.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, et al. 2011. The variant call format and VCFtools. *Bioinformatics* 27:2156-2158.

Darrow, E. M., M. H. Huntley, O. Dudchenko, E. K. Stamenova, N. C. Durand, Z. Sun, S.-C. Huang, A. L. Sanborn, I. Machol, and M. Shamim. 2016. Deletion of DXZ4 on the human inactive X chromosome alters higher-order genome architecture. *Proceedings of the National Academy of Sciences of the United States of America* :201609643.

de Queiroz, K. 1998. The general lineage concept of species, species criteria, and the process of speciation: a conceptual unification and terminological recommendations *in* D. J. Howard, and S. H. Berlocher, eds. Endless Forms: Species and Speciation. Oxford University Press, Oxford, England.

Deakin, J. E., M. J. Edwards, H. Patel, D. O'Meally, J. Lian, R. Stenhouse, S. Ryan, A. M. Livernois, B. Azad, and C. E. Holleley. 2016. Anchoring genome sequence to chromosomes of the central bearded dragon (*Pogona vitticeps*) enables reconstruction of ancestral squamate macrochromosomes and identifies sequence content of the Z chromosome. *BMC Genomics* 17.

Degnan, J. H. and N. A. Rosenberg. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology and Evolution* 24:332-340.

Dixon, J. R., D. U. Gorkin, and B. Ren. 2016. Chromatin domains: the unit of chromosome organization. *Molecular Cell* 62:668-680.

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

Dobzhansky, T. H. 1936. Studies on hybrid sterility. II. Localization of sterility factors in Drosophila pseudoobscura hybrids. *Genetics* 21:113.

Douglas, M. E., M. R. Douglas, G. W. Schuett, L. W. Porras, and A. T. Holycross. 2002. Phylogeography of the western rattlesnake (*Crotalus viridis*) complex, with emphasis on the Colorado Plateau *in* G. W. Schuett, M. Hoggren, M. E. Douglas, and H. W. Greene, eds. Biology of the Vipers. Eagle Mountain Publishing.

Dowell, N. L., M. W. Giorgianni, V. A. Kassner, J. E. Selegue, E. E. Sanchez, and S. B. Carroll. 2016. The Deep Origin and Recent Loss of Venom Toxin Genes in Rattlesnakes. *Current Biology* 26:2434-2445.

Drummond, A. J. and A. Rambaut. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* 7.

Drummond, A. J., A. Rambaut, B. Shapiro, and O. G. Pybus. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution* 22:1185-1192.

Dubey, S., G. P. Brown, T. Madsen, and R. Shine. 2008. Male-biased dispersal in a tropical Australian snake (*Stegonotus cucullatus*, Colubridae). *Molecular Ecology* 17:3506-3514.

Durand, E. Y., N. Patterson, D. Reich, and M. Slatkin. 2011. Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution* 28:2239-2252.

Durand, N. C., M. S. Shamim, I. Machol, S. S. P. Rao, M. H. Huntley, E. S. Lander, and E. L. Aiden. 2016. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Systems* 3:95-98.

Durban, J., A. Perez, L. Sanz, A. Gomez, F. Bonilla, S. Rodriguez, D. Chacon, M. Sasa, Y. Angulo, J. M. Gutierrez, et al. 2013. Integrated "omics" profiling indicates that miRNAs are modulators of the ontogenetic venom composition shift in the Central American rattlesnake, Crotalus simus simus. *BMC Genomics* 14:234.

Duvall, D., S. J. Arnold, and G. W. Schuett. 1992. Pitviper mating systems: ecological potential, sexual selection, and microevolution *in* J. A. Campbell, and E. D. Brodie, eds. Biology of the Pitvipers. Selva, Tyler, TX.

Earl, D. A. and B. M. Vonholdt. 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* 4:359-361.

Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32:1792-1797.

Elith, J., C. H. Graham, R. P. Anderson, M. Dudik, S. Ferrier, A. Guisan, R. J. Hijmans, F. Huettmann, J. R. Leathwick, A. Lehmann, et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography* 29:129-151.

Elith, J., M. Kearney, and S. Phillips. 2010. The art of modelling range-shifting species. *Methods in Ecology and Evolution* 1:330-342.

Ellegren, H., L. Smeds, R. Burri, P. I. Olason, N. Backstrom, T. Kawakami, A. Kunstner, H. Makinen, K. Nadachowska-Brzyska, A. Qvarnstrom, et al. 2012. The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature* 491:756-760.

Ellison, C. K., R. S. Burton, and D. Promislow. 2006. Disruption of mitochondrial function in interpopulation hybrids of *Tigriopus californicus*. *Evolution* 60:1382-1391.

Ellison, C. K., O. Niehuis, and J. Gadau. 2008. Hybrid breakdown and mitochondrial dysfunction in hybrids of *Nasonia* parasitoid wasps. *Journal of Evolutionary Biology* 21:1844-1851.

Evanno, G., S. Regnaut, and J. Goudet. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* 14:2611-2620.

Farallo, V. R. and M. R. J. Forstner. 2012. Predation and the maintenance of color polymorphism in a habitat specialist squamate. *PloS One* 7:e30316.

Faria, R., S. Renaut, J. Galindo, C. Pinho, J. Melo-Ferreira, M. Melo, F. Jones, W. Salzburger, D. Schluter, and R. Butlin. 2014. Advances in Ecological Speciation: an integrative approach. *Molecular Ecology* 23:513-521.

Feder, J. L., S. P. Egan, and P. Nosil. 2012. The genomics of speciation-with-gene-flow. *Trends in Genetics* 28:342-350.

Ferchaud, A. L. and M. M. Hansen. 2016. The impact of selection, gene flow and demographic history on heterogeneous genomic divergence: three-spine sticklebacks in divergent environments. *Molecular Ecology* 25:238-259.

Fishman, L. and J. H. Willis. 2006. A cytonuclear incompatibility causes anther sterility in *Mimulus* hybrids. *Evolution* 60:1372-1381.

Flaxman, S. M., J. L. Feder, and P. Nosil. 2013. Genetic hitchhiking and the dynamic buildup of genomic divergence during speciation with gene flow. *Evolution* 67:2577-2591.

Flores-Villela, O. and E. A. Martinez-Salazar. 2009. Historical explanation of the origin of the herpetofauna of Mexico. *Reviews in Mexican Biodiversity* 80:817-833.
Fu, Y.-X. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* 147:915-925.
Fujita, M. K., S. V. Edwards, and C. P. Ponting. 2011. The *Anolis* lizard genome: an amniote genome without isochores. *Genome Biology and Evolution* 3:974-984.
Gagnaire, P.-A., E. Normandeau, and L. Bernatchez. 2012. Comparative genomics reveals adaptive protein evolution and a possible cytonuclear incompatibility between European and American eels. *Molecular Biology and Evolution* 29:2909-2919.
Gamble, T., T. A. Castoe, S. V. Nielsen, J. L. Banks, D. C. Card, D. R. Schield, G. W. Schuett, and W. Booth. 2017. The Discovery of XY Sex Chromosomes in a Boa and Python. *Current Biology* 27:2148-2153 e2144.
Gans, C. 1961. The feeding mechanism of snakes and its possible evolution. *American Zoologist* :217-227.
Gay, L., P. A. Crochet, D. A. Bell, and T. Lenormand. 2008. Comparing clines on molecular and phenotypic traits in hybrid zones: a window on tension zone models. *Evolution* 62:2789-2806.
Georges, A., Q. Li, J. Lian, D. O'Meally, J. Deakin, Z. Wang, P. Zhang, M. Fujita, H. R. Patel, and C. E. Holleley. 2015. High-coverage sequencing and annotated assembly of the genome of the Australian dragon lizard *Pogona vitticeps*. *GigaScience* 4:45.
Glenn, J. L. and R. Straight. 1977. Mojave Rattlesnake *Crotalus scutulatus scutulatus* venom - variation in toxicity with geographical origin. *Toxicon* 16:81-84.
Glenn, J. L. and R. C. Straight. 1989. Intergradation of 2 Different Venom Populations of the Mojave Rattlesnake (*Crotalus scutulatus scutulatus*) in Arizona. *Toxicon* 27:411-418.
Glenn, J. L., R. C. Straight, M. C. Wolfe, and D. L. Hardy. 1983. Geographical variation in *Crotalus scutulatus scutulatus* (Mojave rattlesnake) venom properties. *Toxicon* 21:119-130.
Gompert, Z. and C. A. Buerkle. 2011. Bayesian estimation of genomic clines. *Molecular Ecology* 20:2111-2127.
Gompert, Z. and C. A. Buerkle. 2012. bgc: Software for Bayesian estimation of genomic clines. *Molecular Ecology Resources* 12:1168-1176.
Gompert, Z., L. K. Lucas, C. C. Nice, J. A. Fordyce, M. L. Forister, and C. A. Buerkle. 2012a. Genomic regions with a history of divergent selection affect fitness of hybrids between two butterfly species. *Evolution* 66:2167-2181.
Gompert, Z., T. L. Parchman, and C. A. Buerkle. 2012b. Genomics of isolation in hybrids. *Philosophical Translations of the Royal Society B* 367:439-450.
Good, J. M., M. A. Handel, and M. W. Nachman. 2008. Asymmetry and polymorphism of hybrid male sterility during the early stages of speciation in house mice. *Evolution* 62:50-65.
Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, and Q. Zeng. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* 29:644.
Graves, J. A. 2016. Evolution of vertebrate sex chromosomes and dosage compensation. *Nature Reviews Genetics* 17:33-46.

Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics* 5:e1000695.
Hannon, G. J. 2014. FASTX-Toolkit.
Hardy, D. L. 1983. Envenomation by the Mojave Rattlesnake (*Crotalus scutulatus scutulatus*) in Southern Arizona, USA. *Toxicon* 21:111-118.
Hargreaves, A. D., M. T. Swain, M. J. Hegarty, D. W. Logan, and J. F. Mulley. 2014. Genomic and transcriptomic insights into the regulation of snake venom production. *bioRxiv*:008474.
Harris, R. S. 2007. Improved pairwise alignment of genomic DNA. The Pennsylvania State University.
Hartl, D. and A. G. Clark. 1997. Principles of population genetics. Sinauer, Sunderland, Massachusetts, USA.
Hey, J. 2010. Isolation with migration models for more than two populations. *Molecular Biology and Evolution* 27:905-920.
Hey, J. and R. Nielsen. 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis. Genetics* 167:747-760.
Hey, J. and R. Nielsen. 2007. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proceedings of the National Academy of Sciences of the United States of America* 104:2785-2790.
Hijmans, R. J., S. E. Cameron, J. L. Parra, P. G. Jones, and A. Jarvis. 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology* 25:1965-1978.
Hillier, L. W., W. Miller, E. Birney, W. Warren, R. C. Hardison, C. P. Ponting, P. Bork, D. W. Burt, M. A. M. Groenen, and M. E. Delany. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432:695-716.
Hoekstra, H. E. 2006. Genetics, development and evolution of adaptive pigmentation in vertebrates. *Heredity* 97:222-234.
Holman, J. A. 1995. Pleistocene amphibians and reptiles of North America. Oxford University Press, New York, NY.
Holman, J. A. 2000. Fossil Snakes of North America: Origion, Evolution, Distribution, Paleoecology. Indiana University Press, Bloomington, IN.
Huang, W., N. Takebayashi, Y. Qi, and M. J. Hickerson. 2011. MTML-msBayes: approximate Bayesian comparative phylogeographic inference from multiple taxa and multiple loci with rate heterogeneity. *BMC Bioinformatics* 12:1.
Hubbard, J. K., J. A. Uy, M. E. Hauber, H. E. Hoekstra, and R. J. Safran. 2010. Vertebrate pigmentation: from underlying genes to adaptive function. *Trends in Genetics* 26:231-239.
Hubisz, M. J., D. Falush, M. Stephens, and J. K. Pritchard. 2009. Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources* 9:1322-1332.
Huelsenbeck, J. P. and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17:754-755.

Ignatiadis, N., B. Klaus, J. B. Zaugg, and W. Huber. 2016. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature Methods* 13:577-580.
Ikeda, N., T. Chijiwa, K. Matsubara, N. Oda-Ueda, S. Hattori, Y. Matsuda, and M. Ohno. 2010. Unique structural characteristics and evolution of a cluster of venom phospholipase A 2 isozyme genes of *Protobothrops flavoviridis* snake. *Gene* 461:15-25.
Irion, U., A. P. Singh, and C. Nusslein-Volhard. 2016. The developmental genetics of vertebrate color pattern formation: lessons from zebrafish. *Current Topics in Developmental Biology* 117:141-169.
Irwin, D. E., M. Alcaide, K. E. Delmore, J. H. Irwin, and G. L. Owens. 2016. Recurrent selection explains parallel evolution of genomic regions of high relative but low absolute differentiation in a ring species. *Molecular Ecology* 25:4488-4507.
Janousek, V., L. Wang, K. Luzynski, P. Dufkova, M. M. Vyskocilova, M. W. Nachman, P. Munclinger, M. Macholan, J. Pialek, and P. K. Tucker. 2012. Genome-wide architecture of reproductive isolation in a naturally occurring hybrid zone between *Mus musculus musculus* and *M. m. domesticus*. *Molecular Ecology* 21:3032-3047.
Jay, F., S. Manel, N. Alvarez, E. Y. Durand, W. Thuiller, R. Holderegger, P. Taberlet, and O. Francois. 2012. Forecasting changes in population genetic structure of alpine plants in response to global warming. *Molecular Ecology* 21:2354-2368.
Jezkova, T., V. Olah-Hemmings, and B. R. Riddle. 2011. Niche shifting in response to warming climate after the last glacial maximum: inference from genetic data and niche assessments in the chisel-toothed kangaroo rat (*Dipodomys microps*). *Global Change Biology* 17:3486-3502.
Jezkova, T., B. R. Riddle, D. C. Card, D. R. Schield, M. E. Eckstut, and T. A. Castoe. 2015. Genetic consequences of postglacial range expansion in two codistributed rodents (genus *Dipodomys*) depend on ecology and genetic locus. *Molecular Ecology* 24:83-97.
Jombart, T., S. Devillard, and F. Balloux. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics* 11:94.
Kass, R. E. and A. E. Raftery. 1995. Bayes Factors. *Journal of the American Statistical Association* 90:773-795.
Keogh, J. S., J. K. Webb, and R. Shine. 2007. Spatial genetic analysis and long-term mark-recapture data demonstrate male-biased dispersal in a snake. *Biology Letters* 3:33-35.
Klauber, L. M. 1930. Differential characteristics of southwestern rattlesnakes allied to *Crotalus atrox*. Zoological Society of San Diego.
Klauber, L. M. 1956. Rattlesnakes: Their Habits, Life Histories, and Influence on Mankind. University of California Press, Berkeley, CA.
Kuehne, H. A., H. A. Murphy, C. A. Francis, and P. D. Sniegowski. 2007. Allopatric divergence, secondary contact, and genetic isolation in wild yeast populations. *Current Biology* 17:407-411.
Kulathinal, R. J., S. M. Bennett, C. L. Fitzpatrick, and M. A. F. Noor. 2008. Fine-scale mapping of recombination rate in *Drosophila* refines its correlation to diversity and divergence. *Proceedings of the National Academy of Sciences of the United States of America* 105:10051-10056.
Kumar, S., G. Stecher, M. Suleski, and S. B. Hedges. 2017. TimeTree: a resource for timelines, timetrees, and divergence times. *Molecular Biology and Evolution* 34:1812-1819.

Lachance, J. and S. A. Tishkoff. 2014. Biased gene conversion skews allele frequencies in human populations, increasing the disease burden of recessive alleles. *The American Journal of Human Genetics* 95:408-420.
Lachumanan, R., A. Armugam, and C.-H. Tan. 1998. Structure and organization of the cardiotoxin genes in Naja naja sputatrix. *FEBS letters* 433:119-124.
Lane, A. and R. Shine. 2011a. Intraspecific variation in the direction and degree of sex-biased dispersal among sea-snake populations. *Molecular Ecology* 20:1870-1876.
Lane, A. and R. Shine. 2011b. Phylogenetic relationships within laticaudine sea snakes (Elapidae). *Molecular Phylogenetics and Evolution* 59:567-577.
Lanfear, R., B. Calcott, S. Y. Ho, and S. Guindon. 2012. Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular Biology and Evolution* 29:1695-1701.
Leache, A. D., M. K. Fujita, V. N. Minin, and R. R. Bouckaert. 2014. Species Delimitation using Genome-Wide SNP Data. *Systematic Biology* 63:534-542.
Leache, A. D., R. B. Harris, M. E. Maliska, and C. W. Linkem. 2013. Comparative species divergence across eight triplets of spiny lizards (*Sceloporus*) using genomic sequence data. *Genome Biology and Evolution* 5:2410-2419.
Lee, C., A. Abdool, and C.-H. Huang. 2009. PCA-based population structure inference with generic clustering algorithms. *BMC Bioinformatics* 10:S73.
Lewontin, R. C. 1972. The apportionment of human diversity *in* T. H. Dobzhansky, M. K. Hecht, and W. C. Steere, eds. Evolutionary Biology.
Lewontin, R. C. and J. Krakauer. 1973. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* 74:175-195.
Li, H. and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754-1760.
Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and S. Genome Project Data Processing. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078-2079.
Li, W. and A. Godzik. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22:1658-1659.
Liao, Y., G. K. Smyth, and W. Shi. 2013. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30:923-930.
Librado, P. and J. Rozas. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25:1451-1452.
Lieberman-Aiden, E., N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326:289-293.
Lin, C.-Y., V. B. Vega, J. S. Thomsen, T. Zhang, S. L. Kong, M. Xie, K. P. Chiu, L. Lipovich, D. H. Barnett, and F. Stossi. 2007. Whole-genome cartography of estrogen receptor α binding sites. *PLoS Genetics* 3:e87.
Lindtke, D., C. A. Buerkle, T. Barbara, B. Heinze, S. Castiglione, D. Bartha, and C. Lexer. 2012. Recombinant hybrids retain heterozygosity at many loci: new insights into the genomics of reproductive isolation in *Populus*. *Molecular Ecology* 21:5042-5058.

Liu, C. R., P. M. Berry, T. P. Dawson, and R. G. Pearson. 2005. Selecting thresholds of occurrence in the prediction of species distributions. *Ecography* 28:385-393.
Liu, L. and L. Yu. 2010. Phybase: an R package for species tree analysis. *Bioinformatics* 26:962-963.
Lowry, D. B., S. Hoban, J. L. Kelley, K. E. Lotterhos, L. K. Reed, M. F. Antolin, and A. Storfer. 2016. Breaking RAD: an evaluation of the utility of restriction site‐associated DNA sequencing for genome scans of adaptation. *Molecular Ecology Resources*.
Luna, M. S. A., T. M. A. Hortencio, Z. S. Ferreira, and N. Yamanouye. 2009. 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. *Journal of Experimental Biology* 212:1535-1543.
Mackessy, S. P. 2008. Venom composition in rattlesnakes: trends and biological significance *in* W. K. Hayes, K. R. Beaman, M. D. Cardwell, and S. P. Bush, eds. The Biology of Rattlesnakes. Loma Linda University Press, Loma Linda, CA.
Mackessy, S. P. 2010. The field of reptile toxinology *in* S. P. Mackessy, ed. Handbook of venoms and toxins of reptiles. CRC Press, New York, NY.
Mahalanobis, P. C. 1936. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)* 2:49-55.
Manni, F., E. Guerard, and E. Heyer. 2004. Geographic patterns of (genetic, morphologic, linguistic) variation: How barriers can be detected by using Monmonier's algorithm. *Human Biology* 76:173-190.
Marçais, G. and C. Kingsford. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27:764-770.
Marshall, C. J. and J. K. Liebherr. 2000. Cladistic biogeography of the Mexican transition zone. *Journal of Biogeography* 27:203-216.
Martin, S. H., K. K. Dasmahapatra, N. J. Nadeau, C. Salazar, J. R. Walters, F. Simpson, M. Blaxter, A. Manica, J. Mallet, and C. D. Jiggins. 2013. Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Research* 23:1817-1828.
Massey, D. J., J. J. Calvete, E. E. Sanchez, L. Sanz, K. Richards, R. Curtis, and K. Boesen. 2012. Venom variability and envenoming severity outcomes of the *Crotalus scutulatus scutulatus* (Mojave rattlesnake) from Southern Arizona. *Journal of Proteomics* 75:2576-2587.
Matsubara, K., H. Tarui, M. Toriba, K. Yamada, C. Nishida-Umehara, K. Agata, and Y. Matsuda. 2006. Evidence for different origin of sex chromosomes in snakes, birds, and mammals and step-wise differentiation of snake sex chromosomes. *Proceedings of the National Academy of Sciences of the United States of America* 103:18190-18195.
Mayr, E. 1942. Systematics and the origin of species, from the viewpoint of a zoologist. Harvard University Press.
Mayr, E. 1963. Animal Species and Evolution. Belknap Press of Harvard University Press, Cambridge, MA.
Meik, J. M., J. W. Streicher, A. M. Lawing, O. Flores-Villela, and M. K. Fujita. 2015. Limitations of climatic data for inferring species boundaries: insights from speckled rattlesnakes. *PloS One* 10:e0131435.

Meik, J. M., J. W. Streicher, E. Mocino-Deloya, K. Setser, and D. Lazcano. 2012. Shallow phylogeographic structure in the declining Mexican lance-headed rattlesnake, *Crotalus polystictus* (Serpentes: Viperidae). *Phyllomedusa* 11:3-11.

Miller, M. P. 2005. Alleles In Space (AIS): Computer software for the joint analysis of interindividual spatial and genetic information. *Journal of Heredity* 96:722-724.

Miller, M. P., M. R. Bellinger, E. D. Forsman, and S. M. Haig. 2006. Effects of historical climate change, habitat connectivity, and vicariance on genetic structure and diversity across the range of the red tree vole (*Phenacomys longicaudus*) in the Pacific Northwestern United States. *Molecular Ecology* 15:145-159.

Miller, S. A., D. D. Dykes, and H. Polesky. 1988. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Research* 16:1215.

Minton, S. A. and S. A. Weinstein. 1986. Geographic and ontogenic variation in venom of the western diamondback rattlesnake (*Crotalus atrox*). *Toxicon* 24:71-80.

Morafka, D. J. 1977a. A biogeographical analysis of the Chihuahuan desert through its herpetofauna. The Hague, Dr. W. Junk BV Publishers.

Morafka, D. J. 1977b. Is there a Chihuahuan Desert? A quantitative evaluation through a herpetofaunal perspective. Pp. 437-454. Transactions of the Symposium on the Biological Resources of the Chihuahuan Desert Region, United States and Mexico. National Park Service, Washington, DC.

Muller, H. J. 1942. Isolating mechanisms, evolution and temperature. Pp. 71-125. Biol. Symp.

Murphy, R. W. and C. Ben Crabtree. 1988. Genetic identification of a natural hybrid rattlesnake: *Crotalus scutulatus scutulatus* × *C. viridis viridis*. *Herpetologica*:119-123.

Murphy, R. W., J. Fu, A. Lathrop, J. V. Feltham, and V. Kovac. 2002. Phylogeny of the rattlesnakes (*Crotalus* and *Sistrurus*) inferred from sequences of five mitochondrial DNA genes *in* G. W. Schuett, M. Hoggren, M. E. Douglas, and H. W. Greene, eds. Biology of the Vipers. Eagle Mountain Publishing.

Myers, E. A., M. J. Hickerson, and F. T. Burbrink. 2016. Asynchronous diversification of snakes in the North American warm deserts. *Journal of Biogeography*.

Myers, S., L. Bottolo, C. Freeman, G. McVean, and P. Donnelly. 2005. A fine-scale map of recombination rates and hotspots across the human genome. *Science* 310:321-324.

Nadalin, F., F. Vezzi, and A. Policriti. 2012. GapFiller: a de novo assembly approach to fill the gap within paired reads. *BMC Bioinformatics* 13:S8.

Nadeau, N. J., A. Whibley, R. T. Jones, J. W. Davey, K. K. Dasmahapatra, S. W. Baxter, M. A. Quail, M. Joron, R. H. ffrench-Constant, M. L. Blaxter, et al. 2012. Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 367:343-353.

Nahas, L., A. S. Kamiguti, M. C. C. S. e Silva, M. A. A. R. de Barros, and P. Morena. 1983. The inactivating effect of *Bothrops jararaca* and *Waglerophis merremii* snake plasma on the coagulant activity of various snake venoms. *Toxicon* 21:239-246.

Nei, M. and W.-H. Li. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America* 76:5269-5273.

Nei, M. and F. Tajima. 1981. DNA Polymorphism Detectable by Restriction Endonucleases. *Genetics* 97:145-163.

Nichol, A. A., V. Douglas, and L. Peck. 1933. On the immunity of rattlesnakes to their venom. *Copeia* 1933:211-213.
Noguchi, H. 1909. Snake Venoms: An investigation of venomous snakes with special reference to the phenomena of their venoms. Carnegie Institution of Washington.
Nosil, P. 2008. Speciation with gene flow could be common. *Molecular Ecology* 17:2103-2106.
Nosil, P., S. P. Egan, and D. J. Funk. 2008. Heterogeneous genomic differentiation between walking-stick ecotypes:"isolation by adaptation" and multiple roles for divergent selection. *Evolution* 62:316-336.
Nosil, P. and J. L. Feder. 2012. Genomic divergence during speciation: causes and consequences Introduction. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 367:332-342.
Nosil, P., D. J. Funk, and D. Ortiz-Barrientos. 2009. Divergent selection and heterogeneous genomic divergence. *Molecular Ecology* 18:375-402.
Nosil, P., T. L. Parchman, J. L. Feder, and Z. Gompert. 2012. Do highly divergent loci reside in genomic regions affecting reproductive isolation? A test using next-generation sequence data in Timema stick insects. *BMC Evolutionary Biology* 12:164.
O'Reilly, P. F., E. Birney, and D. J. Balding. 2008. Confounding between recombination and selection, and the Ped/Pop method for detecting selection. *Genome Research* 18:1304-1313.
Orr, H. A. 1995. The Population-Genetics of Speciation - the Evolution of Hybrid Incompatibilities. *Genetics* 139:1805-1813.
Orr, M. R. and T. B. Smith. 1998. Ecology and speciation. *Trends in Ecology and Evolution* 13:502-506.
Osborne, O. G., T. E. Batstone, S. J. Hiscock, and D. A. Filatov. 2013. Rapid Speciation with Gene Flow Following the Formation of Mt. Etna. *Genome Biology and Evolution* 5:1704-1715.
Oshlack, A., M. D. Robinson, and M. D. Young. 2010. From RNA-seq reads to differential expression results. *Genome Biology* 11:220.
Palumbi, S. R. and C. S. Baker. 1994. Contrasting Population-Structure from Nuclear Intron Sequences and Mtdna of Humpback Whales. *Molecular Biology and Evolution* 11:426-435.
Parchman, T. L., Z. Gompert, M. J. Braun, R. T. Brumfield, D. B. McDonald, J. A. Uy, G. Zhang, E. D. Jarvis, B. A. Schlinger, and C. A. Buerkle. 2013. The genomic consequences of adaptive divergence and reproductive isolation between species of manakins. *Molecular Ecology* 22:3304-3317.
Pasquesi, G. I., R. H. Adams, D. C. Card, D. R. Schield, A. B. Corbin, B. W. Perry, J. Reyes-Velasco, R. P. Ruggiero, M. W. Vandewege, J. A. Shortt, et al. *In Review*. Squamate reptiles challenge paradigms of genomic repeat element evolution set by birds and mammals.
Pawlak, J. and R. M. Kini. 2008. Unique gene organization of colubrid three-finger toxins: complete cDNA and gene sequences of denmotoxin, a bird-specific toxin from colubrid snake *Boiga dendrophila* (Mangrove Catsnake). *Biochimie* 90:868-877.
Payseur, B. A. 2010. Using differential introgression in hybrid zones to identify genomic regions involved in speciation. *Molecular Ecology Resources* 10:806-820.

Payseur, B. A. and L. H. Rieseberg. 2016. A genomic perspective on hybridization and speciation. *Molecular Ecology* 25:2337-2360.
Perez, J. C., S. Pichyangkul, and V. E. Garcia. 1979. The resistance of three species of warm-blooded animals to Western diamondback rattlesnake (*Crotalus atrox*) venom. *Toxicon* 17:601-607.
Perry, B. W., D. C. Card, J. W. McGlothlin, G. I. Pasquesi, N. R. Hales, A. B. Corbin, R. H. Adams, D. R. Schield, M. K. Fujita, J. P. Demuth, et al. *In Review*. Molecular adaptations for sensing and securing prey, and insight into amniote genome diversity, revealed by the garter snake genome.
Peterson, B. K., J. N. Weber, E. H. Kay, H. S. Fisher, and H. E. Hoekstra. 2012. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PloS One* 7:e37135.
Phillips, S. J., R. P. Anderson, and R. E. Schapire. 2006. Maximum entropy modeling of species geographic distributions. *Ecological Modeling* 190:231-259.
Pinho, C. and J. Hey. 2010. Divergence with Gene Flow: Models and Data. *Annual Reviews in Ecology and Evolution* 41:215-230.
Platt, R. N., L. Blanco-Berdugo, and D. A. Ray. 2016. Accurate transposable element annotation is vital when analyzing new genome assemblies. *Genome Biology and Evolution* 8:403-410.
Portik, D., A. D. Leaché, D. Rivera, M. Barej, M. Hirschfeld, M. Burger, M. Rödel, D. Blackburn, and M. K. Fujita. *In Review*. Evaluating mechanisms of diversification in a Guineo-Congolian forest frog using demographic model selection. Molecular Ecology.
Presgraves, D. C. 2010. The molecular evolutionary basis of species formation. *Nature Reviews Genetics* 11:175-180.
Pritchard, J. K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155:945-959.
Puritz, J. B., C. M. Hollenbeck, and J. R. Gold. 2014. dDocent: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms. *PeerJ* 2:e431.
Qualls, C. P. 1997. The effects of reproductive mode and climate on reproductive success in the Australian lizard, *Lerista bougainvillii*. *Journal of Herpetology* :60-65.
Quinlan, A. R. and I. M. Hall. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841-842.
R Core Team. 2017. R: A language and environment for statistical computing.
Rádis-Baptista, G., T. Kubo, N. Oguiura, M. Svartman, T. M. B. Almeida, R. F. Batistic, E. B. Oliveira, Â. M. Vianna-Morgante, and T. Yamane. 2003. Structure and chromosomal localization of the gene for crotamine, a toxin from the South American rattlesnake, *Crotalus durissus terrificus*. *Toxicon* 42:747-752.
Raes, N. and H. ter Steege. 2007. A null-model for significance testing of presence-only species distribution models. *Ecography* 30:727-736.
Ramírez, F., V. Bhardwaj, L. Arrigoni, K. C. Lam, B. A. Grüning, J. Villaveces, B. Habermann, A. Akhtar, and T. Manke. 2018. High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nature Communications* 9:189.
Rao, S. S. P., M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, and E. S. Lander. 2014. A 3D map of the human

genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159:1665-1680.

Reyes-Velasco, J., D. C. Card, A. L. Andrew, K. J. Shaney, R. H. Adams, D. R. Schield, N. R. Casewell, S. P. Mackessy, and T. A. Castoe. 2015. Expression of venom gene homologs in diverse python tissues suggests a new model for the evolution of snake venom. *Molecular Biology and Evolution* 32:173-183.

Reyes-Velasco, J., J. M. Meik, E. N. Smith, and T. A. Castoe. 2013. Phylogenetic relationships of the enigmatic longtailed rattlesnakes (*Crotalus ericsmithi, C. lannomi,* and *C. stejnegeri*). *Molecular Phylogenetics and Evolution* 69:524-534.

Rice, E. S., S. Kohno, J. S. John, S. Pham, J. Howard, L. F. Lareau, B. L. O'Connell, G. Hickey, J. Armstrong, A. Deran, et al. 2017. Improved genome assembly of American alligator genome reveals conserved architecture of estrogen signaling. *Genome Research* 27:686-696.

Robinson, M. D., D. J. McCarthy, and G. K. Smyth. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26:139-140.

Rosenberg, N. A. 2004. DISTRUCT: a program for the graphical display of population structure. *Molecular Ecology Notes* 4:137-138.

Rosenblum, E. B. 2006. Convergent evolution and divergent selection: lizards at the White Sands ecotone. *The American Naturalist* 167:1-15.

Sambatti, J., D. Ortiz‐Barrientos, E. J. Baack, and L. H. Rieseberg. 2008. Ecological selection maintains cytonuclear incompatibilities in hybridizing sunflowers. *Ecology Letters* 11:1082-1091.

Sanchez, E. E., J. A. Galan, R. L. Powell, S. R. Reyes, J. G. Soto, W. K. Russell, D. H. Russell, and J. C. Perez. 2005. Disintegrin, hemorrhagic, and proteolytic activities of Mohave rattlesnake, *Crotalus scutulatus scutulatus* venoms lacking Mojave toxin. *Comparative Biochemistry and Physiology - part C; Toxinology* 141:124-132.

Schield, D. R., R. H. Adams, D. C. Card, B. W. Perry, G. M. Pasquesi, T. Jezkova, D. M. Portik, A. L. Andrew, C. L. Spencer, and E. E. Sanchez. 2017. Insight into the roles of selection in speciation from genomic patterns of divergence and introgression in secondary contact in venomous rattlesnakes. *Ecology and Evolution* 7: 3951-3966.

Schield, D. R., D. C. Card, R. H. Adams, T. Jezkova, J. Reyes-Velasco, F. N. Proctor, C. L. Spencer, H. W. Herrmann, S. P. Mackessy, and T. A. Castoe. 2015. Incipient speciation with biased gene flow between two lineages of the Western Diamondback Rattlesnake (*Crotalus atrox*). *Molecular Phylogenetics and Evolution* 83:213-223.

Schluter, D. and G. L. Conte. 2009. Genetics and ecological speciation. *Proceedings of the National Academy of Sciences of the United States of America* 106:9955-9962.

Schuett, G. W., R. A. Repp, M. Amarello, and C. F. Smith. 2013. Unlike most vipers, female rattlesnakes (*Crotalus atrox*) continue to hunt and feed throughout pregnancy. *Journal of Zoology* 289:101-110.

Secor, S. and J. Diamond. 1998. A vertebrate model of extreme physiological regulation. *Nature* 395:659-662.

Simão, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31:3210-3212.

Simpson, G. G. 1951. The Species Concept. *Evolution* 5:285-298.
Slatkin, M. and L. Excoffier. 2012. Serial founder effects during range expansion: a spatial analog of genetic drift. *Genetics* 191:171-181.
Sloan, D. B., J. C. Havird, and J. Sharbrough. 2016. The on‐again, off‐again relationship between mitochondrial genomes and species boundaries. *Molecular Ecology* 26: 2212-2236.
Smeds, L., T. Kawakami, R. Burri, P. Bolivar, A. Husby, A. Qvarnstrom, S. Uebbing, and H. Ellegren. 2014. Genomic identification and characterization of the pseudoautosomal region in highly differentiated avian sex chromosomes. *Nature Communications* 5:5448.
Smit, A. F. and R. Hubley. 2015. RepeatModeler Open 1.0.
Smit, A. F. A., R. Hubley, and P. Green. 2015. RepeatMasker Open-4.0. 2013–2015. Institute for Systems Biology. http://repeatmasker. org.
Smith, C. F. and S. P. Mackessy. 2016. The effects of hybridization on divergent venom phenotypes: Characterization of venom from *Crotalus scutulatus scutulatus* x *Crotalus oreganus helleri* hybrids. *Toxicon* 120:110-123.
Solovyev, V., P. Kosarev, I. Seledsov, and D. Vorobyev. 2006. Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biology* 7:S10.
Spencer, C. L. 2003. Geographic variation in morphology, diet, and reproduction of a widespread pitviper, the Western Diamondback Rattlesnake (*Crotalus atrox*). Department of Biology. University of Texas at Arlington, Arlington, TX.
Spencer, C. L. 2008. Geographic variation in western diamond-backed rattlesnake (*Crotalus atrox*) morphology. Pp. 55-78 *in* W. K. Hayes, K. R. Beaman, M. D. Cardwell, and S. P. Bush, eds. *The Biology of Rattlesnakes*. Loma Linda University Press, Loma Linda, CA.
Srikulnath, K., C. Nishida, K. Matsubara, Y. Uno, A. Thongpan, S. Suputtitada, S. Apisitwanich, and Y. Matsuda. 2009. Karyotypic evolution in squamate reptiles: comparative gene mapping revealed highly conserved linkage homology between the butterfly lizard (*Leiolepis reevesii rubritaeniata*, Agamidae, Lacertilia) and the Japanese four-striped rat snake (*Elaphe quadrivirgata*, Colubridae, Serpentes). *Chromosome Research* 17:975-986.
Stanke, M. and B. Morgenstern. 2005. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Research* 33:W465-W467.
Streicher, J. W., T. J. Devitt, C. S. Goldberg, J. H. Malone, H. Blackmon, and M. K. Fujita. 2014. Diversification and asymmetrical gene flow across time and space: lineage sorting and hybridization in polytypic barking frogs. *Molecular Ecology* 23:3273-3291.
Streicher, J. W., J. P. McEntee, L. C. Drzich, D. C. Card, D. R. Schield, U. Smart, C. L. Parkinson, T. Jezkova, E. N. Smith, and T. A. Castoe. 2016. Genetic surfing, not allopatric divergence, explains spatial sorting of mitochondrial haplotypes in venomous coralsnakes. *Evolution* 70:1435-1449.
Sweet, S. S. 1985. Geographic variation, convergent crypsis and mimicry in gopher snakes (Pituophis melanoleucus) and western rattlesnakes (*Crotalus viridis*). *Journal of Herpetology* 19:55-67.
Taylor, E. B., J. W. Boughman, M. Groenenboom, M. Sniatynski, D. Schluter, and J. L. Gow. 2006. Speciation in reverse: morphological and genetic evidence of the collapse of a three-spined stickleback (*Gasterosteus aculeatus*) species pair. *Molecular Ecology* 15:343-355.

Toews, D. P. L. and A. Brelsford. 2012. The biogeography of mitochondrial and nuclear discordance in animals. *Molecular Ecology* 21:3907-3930.
Trier, C. N., J. S. Hermansen, G.-P. Sætre, and R. I. Bailey. 2014. Evidence for mito-nuclear and sex-linked reproductive barriers between the hybrid Italian sparrow and its parent species. *PLoS Genetics* 10:e1004075.
Uebbing, S., A. Konzer, L. Xu, N. Backström, B. Brunström, J. Bergquist, and H. Ellegren. 2015. Quantitative mass spectrometry reveals partial translational regulation for dosage compensation in chicken. *Molecular Biology and Evolution* 32:2716-2725.
Ulloa, M., J. N. Corgan, and M. Dunford. 1995. Evidence for nuclear-cytoplasmic incompatibility between *Allium fistulosum* and *Allium cepa*. *Theoretical and Applied Genetics* 90:746-754.
UniProt, C. 2017. UniProt: the universal protein knowledgebase. *Nucleic Acids Research* 45:D158-D169.
Verity, R., C. Collins, D. C. Card, S. M. Schaal, L. Wang, and K. E. Lotterhos. 2016. minotaur: A platform for the analysis and visualization of multivariate results from genome scans with R Shiny. *Molecular Ecology Resources* 7:33-43.
Vicoso, B., J. J. Emerson, Y. Zektser, S. Mahajan, and D. Bachtrog. 2013. Comparative sex chromosome genomics in snakes: differentiation, evolutionary strata, and lack of global dosage compensation. *PLoS Biology* 11:e1001643.
Vonk, F. J., N. R. Casewell, C. V. Henkel, A. M. Heimberg, H. J. Jansen, R. J. R. McCleary, H. M. E. Kerkkamp, R. A. Vos, I. Guerreiro, J. J. Calvete, et al. 2013. The king cobra genome reveals dynamic gene evolution and adaptation in the snake venom system. *Proceedings of the National Academy of Sciences of the United States of America* 110:20651-20656.
Waltari, E., R. J. Hijmans, A. T. Peterson, A. S. Nyari, S. L. Perkins, and R. P. Guralnick. 2007. Locating Pleistocene Refugia: Comparing Phylogeographic and Ecological Niche Model Predictions. *PloS One* 2.
Warren, W. C., D. F. Clayton, H. Ellegren, A. P. Arnold, L. W. Hillier, A. Kunstner, S. Searle, S. White, A. J. Vilella, S. Fairley, et al. 2010. The genome of a songbird. *Nature* 464:757-762.
Watson, D. F. and G. M. Philip. 1985. A Refinement of Inverse Distance Weighted Interpolation. *Geo-Processing* 2:315-327.
Webb, W. C., J. M. Marzluff, and K. E. Omland. 2011. Random interbreeding between cryptic lineages of the Common Raven: evidence for speciation in reverse. *Molecular Ecology* 20:2390-2402.
Weir, B. S. and C. C. Cockerham. 1984. Estimating F-statistics for the analysis of population structure. *Evolution* 38:1358-1370.
Weirauch, M. T., A. Yang, M. Albu, A. G. Cote, A. Montenegro-Montero, P. Drewe, H. S. Najafabadi, S. A. Lambert, I. Mann, and K. Cook. 2014. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* 158:1431-1443.
Wiens, J. J., T. N. Engstrom, and P. T. Chippindale. 2006. Rapid diversification, incomplete isolation, and the "speciation clock" in North American salamanders (genus *Plethodon*): testing the hybrid swarm hypothesis of rapid radiation. *Evolution* 60:2585-2603.
Wiley, E. O. 1978. Evolutionary Species Concept Reconsidered. *Systematic Zoology* 27:17-26.

Wuster, W., M. da Graca Salomao, J. A. Quijada-Mascarenas, R. S. Thorpe, and B.-B. B. S. Project. 2002. Origin and evolution of the South American pitviper fauna: evidence from mitochondrial DNA sequence data *in* G. W. Schuett, M. Hoggren, M. E. Douglas, and H. Green, eds. *Biology of the Vipers*. Eagle Mountain Publishing, Salt Lake City, UT.

Xue, W., J.-T. Li, Y.-P. Zhu, G.-Y. Hou, X.-F. Kong, Y.-Y. Kuang, and X.-W. Sun. 2013. L_RNA_scaffolder: scaffolding genomes with transcripts. *BMC Genomics* 14:604.

Yang, Z. and B. Rannala. 2010. Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences of the United States of America* 107:9264-9269.

Yin, W., Z. J. Wang, Q. Y. Li, J. M. Lian, Y. Zhou, B. Z. Lu, L. J. Jin, P. X. Qiu, P. Zhang, W. B. Zhu, et al. 2016. Evolutionary trajectories of snake genes and genomes revealed by comparative analyses of five-pacer viper. *Nature communications* 7:13107.

Zancolli, G., T. G. Baker, A. Barlow, R. K. Bradley, J. J. Calvete, K. C. Carter, K. de Jager, J. B. Owens, J. F. Price, and L. Sanz. 2016. Is hybridization a source of adaptive venom variation in rattlesnakes? a test, using a *Crotalus scutulatus* × *viridis* hybrid zone in southwestern New Mexico. *Toxins* 8:188.

Zhang, C., D. X. Zhang, T. Zhu, and Z. Yang. 2011. Evaluation of a bayesian coalescent method of species delimitation. *Systematic Biology* 60:747-761.

Zhao, N. L. a. H. 2006. A non-parametric approach to population structure inference using multilocus genotypes. *Human Genomics* 2:353.