POPULATION GENETIC AND GENOMIC ANALYSES OF WESTERN MASSASAUGA (*SISTRURUS TERGEMINUS* SSP.): SUBSPECIES DELIMITATION AND CONSERVATION STATUS

by

Rian Bylsma

A Thesis

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Master of Science
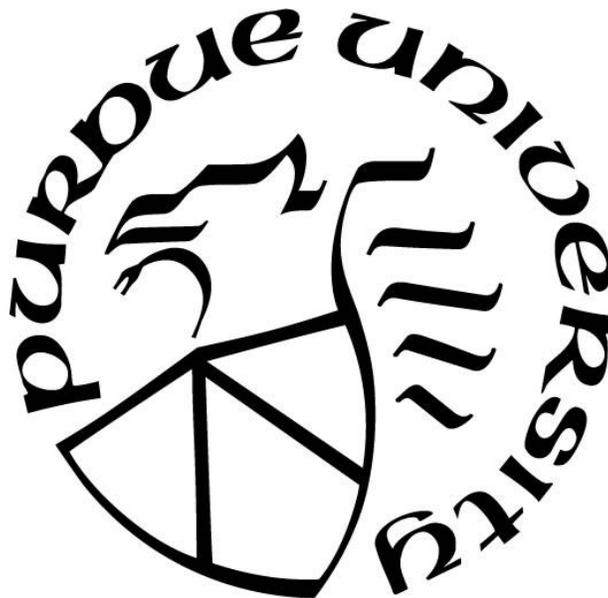
Department of Forestry and Natural Resources

West Lafayette, Indiana

December 2020

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. J. Andrew DeWoody, Chair

Department of Forestry & Natural Resources; Department of Biological Sciences

Dr. Ximena Bernal

Department of Biological Sciences

Dr. Anna Brüniche-Olsen

Department of Forestry & Natural Resources

Dr. Lisle Gibbs

Department of Evolution, Ecology, and Organismal Biology, The Ohio State University

Dr. Rod Williams

Department of Forestry & Natural Resources

Approved by:

Dr. Robert Wagner

Dedicated to my family and friends for their patient support

ACKNOWLEDGMENTS

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The Western Massasauga (*Sistrurus tergeminus*) is a small, North American rattlesnake found west of the Mississippi River. *Sistrurus tergeminus* has previously been divided into two putative subspecies, Desert (*S. t. edwardsii*) and Prairie Massasaugas (*S. t. tergeminus*) based upon qualitative variation in morphology, coloration, and habitat. The Desert Massasauga subspecies has been formally petitioned for federal listing under the U.S. Endangered Species Act. Our overarching goal was to evaluate genetic structure and genomic differentiation between specimens of the two putative subspecies in an effort to inform ongoing conservation assessments. To that end, we generated whole genome sequence data for both putative taxa and then developed nearly 200 genetic markers from different fractions of the genome (~50 intergenic and ~50 genic markers from each of the two subspecies) to test for population structure across much of the Western Massasauga range. Mean genomic divergence between subspecies was only 0.0041 ± 0.0080 (Kimura's 2-parameter distance) for nuclear sequences and 0.0175 ± 0.0031 for mitochondrial sequences, both exceedingly low values which approach the minimum of zero. Admixture analyses and F-statistics both indicated that regardless of how the markers were partitioned, genetic structure was oriented far more along a geographic axis (isolation-by-distance) than a taxonomic axis (i.e., between putative subspecies). Overall, our analyses provide little support that formal protection of the purported Desert Massasauga is warranted based on the homogeneity of the collective Western Massasauga gene pool.

CHAPTER 1.   POPULATION GENETIC AND GENOMIC ANALYSES OF WESTERN MASSASAUGA (*SISTRURUS TERGEMINUS SSP.*): SUBSPECIES DELIMITATION AND CONSERVATION STATUS

A copy of this work is in review under the same title, Population genetic and genomic analyses of Western Massasauga (*Sistrurus tergeminus ssp.*): subspecies delimitation and conservation status.

1.1   Introduction

Rattlesnake populations are subject to a long litany of existential threats including climate change, disease, overexploitation, human persecution, habitat degradation and fragmentation (Clark et al. 2010; Clark et al. 2011; Colley et al. 2017; Fitzgerald and Painter 2000). Many rattlesnake populations are now of conservation concern (at least to biologists, if not to much of the general public), but legal protections generally require biological delineation. For example, the U.S. Endangered Species Act (ESA) requires that listings apply to "distinct population segments." Ideally, these biotic entities coincide with the formal, accepted Linnean taxonomy, but unfortunately that is not always the case (O'Brien and Mayr 1991).

The taxonomy of rattlesnakes is convoluted and has been controversial at times (Crother et al. 2011, 2013; Holycross et al. 2008; Knight et al. 1993; Murphy et al. 2002). Today, rattlesnakes are generally considered a monophyletic lineage of American pit vipers that are divided into two reciprocally monophyletic genera, *Crotalus* and *Sistrurus*, that diverged from one another about 12.5 MYA (Figure 1; Blair and Sánchez-Ramírez 2016). Much of the phylogenetic research on *Sistrurus* sp. has been motivated by research on the Eastern Massasauga (*S. catenatus*). Notably, DNA sequences have confirmed the evolutionary distinctiveness of *S. catenatus* and have clarified the evolutionary taxonomy of the Pygmy Rattlesnake (*S. miliarius*) and the Western Massasauga (*S. tergeminus*), but they provide relatively weak support for distinguishing *S. tergeminus edwardsii* (Desert Massasauga) from *S. tergeminus tergeminus* (Prairie Massasauga) (Crother et al. 2011; Kubatko et al. 2011). The few available sequence data suggest that genomic differentiation (if any) between these two putative taxa may be limited due to various interacting factors (e.g., gene flow, overlapping ranges, etc.).
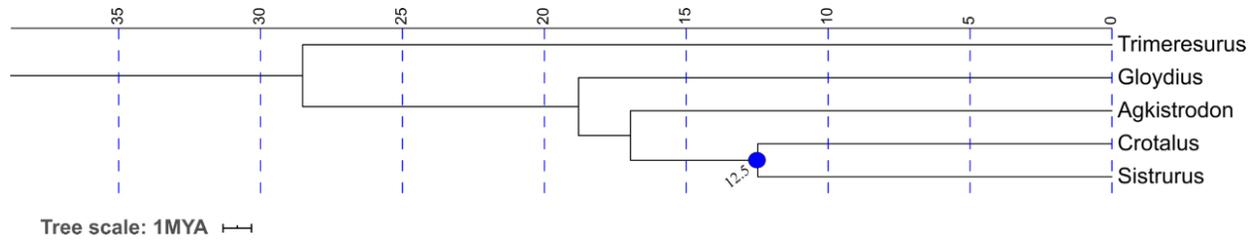
Figure 1.1. Evolutionary tree of the clade in Viperidae containing *Sistrurus*. The two rattlesnake genera (*Sistrurus* and *Crotalus*) diverged from one another about 12.5 MYA (TimeTree.org; Kumar et al. 2017)

Western Massasauga subspecies could be in the early stages of ecological speciation (Kubatko et al. 2011). The Desert Massasauga, historically recognized in the American Southwest from northern Mexico to southern Colorado and central Texas, is a small rattlesnake that lives in xeric grasslands feeding largely off centipedes and lizards whereas the Prairie Massasauga, found in central Texas up to central Iowa and Missouri, occupies mesic grasslands and feeds more on small mammals (Holycross and Mackessy 2002). The original distinction between these two subspecies is traced back to surveys by Howard Gloyd based on the Prairie Massasauga's larger size, darker ventral coloration, higher number of midbody scale rows, and more numerous ventral scales and dorsal blotches (Gloyd 1955) but there is no obvious geographical barrier separating Desert and Prairie Massasaugas (Ryberg et al. 2015; Figure 2). Genetic diversity has been studied in specific populations (Anderson et al. 2009; Gibbs et al. 2011; McCluskey and Bender 2015), but has not definitively resolved the issue of whether the Western Massasauga consists of one or two distinct gene pools (Kubatko et al. 2011; Ryberg et al. 2015).

Our purpose herein is not to reevaluate the phylogenetics or systematics of rattlesnakes but to evaluate the contiguity (or lack thereof) in the Western Massasauga gene pool to help evaluate claims that the Desert Massasauga (*S. t. edwardsii*) should be afforded formal conservation protection (WildEarth Guardians 2010). By virtue of the legal language in the ESA (i.e., "distinct population segment"), it is important to know if the two subspecies of Western Massasaugas have differentiated gene pools or whether they are genetically homogenous. Ryberg et al. (2015) found that genetic structure was inconsistent with subspecies designations in the Western Massasauga, but they only evaluated a small number of DNA sequences and a more expansive survey is desirable.

10

Our first goal was to sequence the genomes of both *S. t. tergeminus* and *S. t. edwardsii* and subsequently generate a curated suite of single nucleotide polymorphism (SNP) markers. We did so by developing a marker array designed a) to best utilize suboptimal DNA obtained from roadkill (our primary source of tissue); b) to avoid ascertainment biases due to potential subspecific differentiation; and c) to query two orthogonal aspects of genetic diversity: functional protein-coding genes as well as more neutral variants in intergenic regions far from known genes. Our second goal was to employ the markers and genomic sequences to ascertain genetic and genomic differentiation between the two putative entities across much of their respective geographic ranges in order to inform impending conservation decisions, with particular emphasis on the use of genomic data in subspecies delineation.

1.2 Materials and Methods

1.2.1 Sample Collection and DNA Extraction

Western Massasauga samples were collected opportunistically via driving surveys (Table A.1). Most snakes were found dead on the road. For the few specimens caught live, we obtained scale clips or shed skins. Three samples from Matagorda Island in southern Texas were recovered as fire mortalities. Additional samples were obtained from previous studies or harvested from museum samples. A map of genotyped sample locations is presented in Figure 2. Both Fetzner's technique (Fetzner 1999) for shed skin and a standard phenol-chloroform extraction for other tissues were used to extract DNA for analysis. Extracted DNA was cleaned using a Zymo dsDNA Clean and Concentrate kit and electrophoresed on agarose gels to confirm DNA quality and quantity. For genome sequencing, one reference sample was chosen from each presumptive species (TJH 3595 and CSA 1). We chose individuals based on DNA quality and quantity as well as on geographic region of origin. Desert Massasauga occur in southern and western Texas whereas Prairie Massasaugas are found in northern and eastern Texas (Mackessy et al. 2005). Therefore, out of the highest quality and quantity DNA samples, we chose samples from Hood and Ward Counties as exemplars within the range of each putative subspecies (Figure 2). We did so to increase the likelihood that each sequenced individual was a representative member of each putative subspecies (see Table A.1 for sample metadata). We confirmed that both samples were

identified in the field as members of the subspecies they are expected to represent based on the amount of dark ventral pigment and number of dorsal scale rows at midbody.
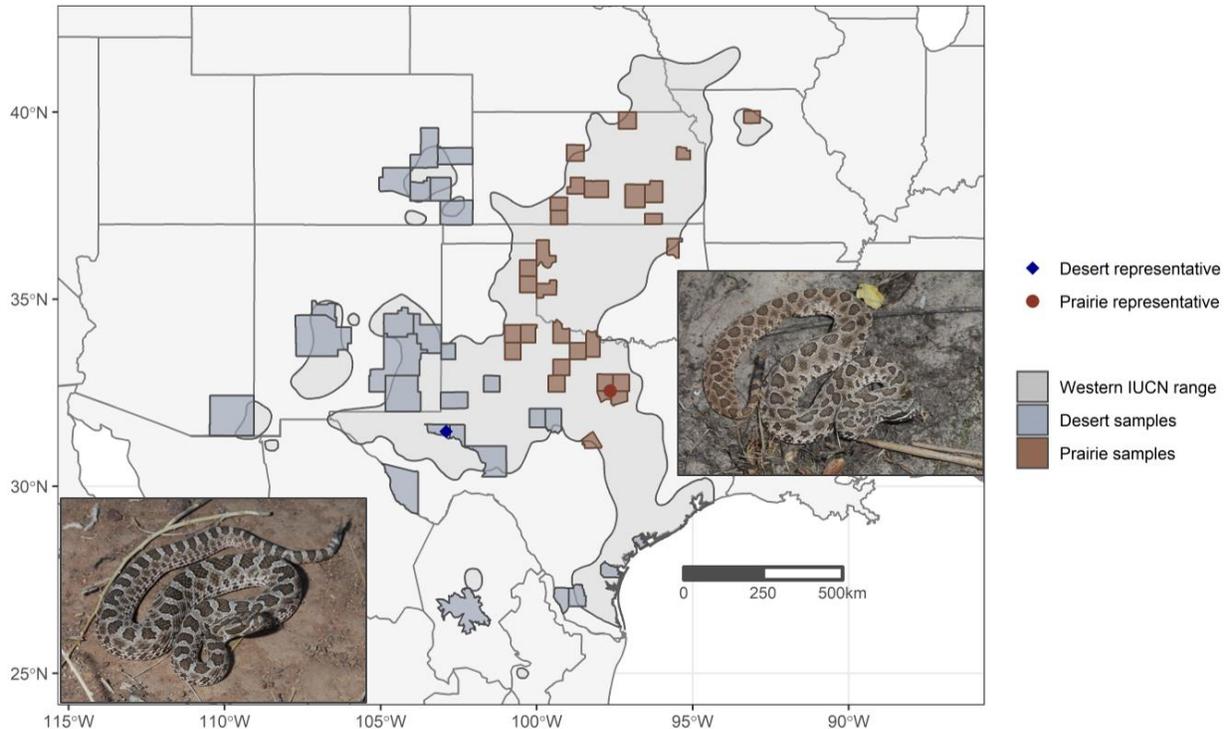
Figure 1.2. Map of sampling locations. The counties in which genotyped individuals were collected are shaded blue for the Desert Massasauga and brown for the Prairie Massasauga. Approximate range of the Western Massasauga is shaded in gray. Note that actual Western Massasauga habitat is highly fragmented throughout much of its range (Anderson et al. 2009). Photos of individuals from Presidio County, Texas (Desert Massasauga; left) and Wheeler County, Texas (Prairie Massasauga; right) are two examples of the Western Massasauga

1.2.2   Genome Sequencing and Assembly

We created independent genomic libraries for each sample, then generated both paired-end and mate-pair reads using an Illumina NovaSeq (S4 2x150) platform. Paired-end library insert sizes were ~350bp and mate-pair insert sizes were restrictred to 300-1500bp. Reads were trimmed with Trimmomatic v0.39 (Bolger et al. 2014) to exclude base qualities less than Phred-20 and read lengths <30bp, and remaining reads were assessed with FastQC v0.11.7 (Andrews 2010). We estimated genome size with kmergenie v1.6982 (Chickhi and Medvedev 2014) and assembled

genomes *de novo* with AByss v2.1.5 (Jackman et al. 2017) and SOAPdenovo2 v240 (Luo et al. 2012).

We also generated mitochondrial genome assemblies for both nominal subspecies. We employed MITObim v1.8 (Hahn et al. 2013), matching kmers of 21bp and syncing paired read data to extend from seed sequences obtained from publicly available data (see Table A.2).

1.2.3 Genomic divergence between putative *S. tergeminus* subspecies

To estimate pairwise nuclear genetic distances between putative taxa, we employed Kimura's 2-parameter (K2P; Kimura 1980) distances using sequences from large scaffolds that are ostensibly orthologous. We did so using scaffolds greater than 100kbp in length from our best assemblies in each subspecies, the *C. viridis* reference genome, and an unpublished assembly for the Eastern Massasauga (courtesy of L. Gibbs; see acknowledgements and online resources). That is, we estimated K2P for all pairwise comparisons of >100kb scaffolds between each pair of taxa. We aligned all genome assemblies pairwise using BLAST+ v2.10.0 (Camacho et al 2009) and calculated K2P distances with the *dist.dna* function of the ape v5.4 R package (Paradis et al. 2004). In Western Massasaugas, these data included 858 scaffolds totaling 0.12 Gb for the Prairie Massasauga and 147 scaffolds comprised of 0.02 Gb for the Desert Massasauga (Table A.3). For each comparison, we plotted the distribution of K2P distances for each aligned sequence with ggplot2 v.3.32 (Wickham 2016). For context, we compiled K2P values from the literature for genetic distance between other (i.e. non-Massasauga) snakes and non-avian reptiles or calculated them ourselves using publicly available genomes and the same methods above.

To estimate pairwise genomic mtDNA distances, we obtained mitochondrial genome assemblies for *Crotalus* (2 spp.) and *Sistrurus* (1 sp.) (see online resources) for comparison with our assemblies. Sequences were then rotated via gapless alignment by CSA (Fernandes et al. 2009) and multiply aligned in MEGA-X v10.1.8 (Kumar et al. 2018) using the ClustalW algorithm (Thompson et al. 1994), and we then estimated pairwise genetic distances again using the K2P model (Kimura 1980). Standard errors for each pairwise distance were estimated with 500 bootstrapped replicates. We again compiled values comparing non-Massasauga snake and non-avian reptile from available mitochondrial datasets.

1.2.4 SNP identification and genotyping

We developed a SNP genotyping array for use in a Fluidigm microfluidic system that works well with nontraditional sources of tissue (e.g., roadkill) that yield suboptimal DNA (Carroll et al. 2018). To identify variants (i.e., candidate SNP markers), we mapped paired end reads against the *C. viridis* reference genome (Schield et al. 2019) because of its chromosome-level assembly, high quality annotation, and the relatively recent divergence time from *S. tergeminus* of 12.5 MYA (Blair and Sánchez-Ramírez 2016). We mapped reads with BWA v0.7.17 (Li 2013) and used Picard v2.18.2 (Broad Institute 2018) to filter and quality check our alignments. We then called variants for each sample using GATK HaplotypeCaller v3.8.1 (Van der Auwera et al. 2013). We hard-filtered variants to include only those with a) read depth between 10x and 100x; b) a strand odds ratio greater than 4.0; c) quality by depth (an estimate of base quality as a function of allele depth) less than 5.0; and d) quality scores greater 30.0. We used IGV v2.5.3 to further restrict these variants to SNPs greater than 20bp from neighboring SNPs and a neighboring GC content less than 65%. Additionally, we chose only SNPs at loci with flanking regions of contiguously mapping reads for at least 10kbp to reduce chances of linkage disequilibrium (LD) between our markers.

From the filtered variants, we designed 96 putatively neutral markers from intergenic regions and 96 putatively "adaptive" markers from genic regions (as indicated by the annotations of the reference genome). To minimize ascertainment bias, we selected half of each marker type listed above (intergenic versus genic) from each representative subspecies, irrespective of whether SNPs were polymorphic in the other subspecies (Figure B.1). Using all 192 loci and the Fluidigm microfluidics platform, we genotyped 88 samples (Table A.1) that represented snakes collected from across the Western Massasauga range (Figure 2). We also included three technical replicates each for one Desert and one Prairie sample to estimate genotyping error rates. Samples were genotyped using the Fluidigm Biomark HD platform.

1.2.5 Population Genetic Analyses

We used the Fluidigm analysis software to call genotypes, pruning data by removing loci that failed to produce distinct genotype clusters in greater than 20% of individuals and by removing individuals that were successfully genotyped at fewer than 80% of remaining loci. We estimated the genotyping error rate (e) from our three technical replicates as described in Doyle et al. (2016)

using the equation $e = \frac{m}{ds}$, where m is the number of pairwise mismatches between replicates of the majority consensus including miscalls and amplification failures, d is the number loci in each replicate, and s is the total number of replicates. We tested for LD between pairs of loci using Genepop v4.7.3 (Rousset 2008) and subsequently removed one member of each linked pair from subsequent analyses.

We computed indices of genetic diversity using Genepop. We calculated minor allele frequency (MAF), observed heterozygosity (H_O), tested for deviations from Hardy-Weinberg Equilibrium (HWE) at each locus, and estimated F_{ST} using various subsets of the 192 markers (ALL, HWE, GENIC, INTERGENIC, STT, and STE) to partition variances in allele frequencies both functionally and taxonomically. Putative subspecies were evaluated as separate datasets and artificially split in half to produce random "populations" for comparison. F_{ST} was then calculated for each of 500 randomization trials using all markers and compared against the Desert-Prairie F_{ST} to discern if the inter-subspecific was greater than the intra-subspecific F_{ST}. We then calculated the F_{ST}, F_{IS}, and r^2 (as a measure of LD) for each locus or pair and the means for all Western Massasauga samples together. Additionally, we adjusted the p-values for deviations from HWE with the sequential Bonferroni method (Waples and Allendorf 2015) across all loci in order to identify locus-specific deviations. To test for the Wahlund effect (overall heterozygote deficiency due to population structure), we plotted the relationships of F_{IS} and F_{ST} for the entire dataset (all individuals of both presumptive subspecies) and calculated a linear regression. If population structure is causing deviations from HWE, we would expect a positive mean F_{IS} and a positive linear relationship with slope = 1 for F_{IS}/F_{ST} (Waples and Allendorf 2015).

We tested for population structure by first conducting a PCA with the 139 HWE loci (i.e., those in HWE) in adegenet v2.1.2 (Jombart 2008). Second, we conducted admixture analyses using the program LEA v2.0 (Frichot and François 2015). For LEA analyses of all data subsets, we included loci regardless of whether they were in HWE and used LEA's minimal cross-entropy approach to determine optimum α and k. Additionally, we ran k = 2 with α = 5 to directly test for signatures of two distinct subspecies.

To test for isolation by distance (IBD), we conducted a Monte-Carlo based Mantel test of genetic distances versus geographic distances using adegenet. For this test, we limited our pairwise distance measurements to samples with geographic coordinates (n=37 individuals, 666 pairwise distances).

1.3 Results

1.3.1 Genome Sequencing and Assembly

We sequenced about 100 Gb per subspecies, averaging approximately 50x coverage for the Desert Massasauga and 46x coverage for the Prairie Massasauga (Tables S4 and S5 for summary statistics). Species in the family Viperidae have a mean genome size of 2.05 ± 0.49 Gb (Gregory 2020) similar to our assembly size estimates for Western Massasauga (1.897 Gb for Desert and 1.937 Gb for Prairie). GC content was very similar for Desert (40.0%) and Prairie (39.75%) Massasaugas. Mitochondrial assemblies were 17,416 bp and 17,396 bp for the Desert and Prairie individuals respectively. From a sample of n=2 indviduals, we can't say whether the length variation (20 bp) is partitioned among subspecies or individuals but we strongly suspect there is individual variation based on the Ryberg et al. (2015) paper.

1.3.2 Genomic divergence between putative *S. tergeminus* subspecies

Our categorical pairwise alignments of large nuclear scaffolds for K2P estimates covered up to 10% of the *S. tergeminus* genome and thus should reasonably reflect the overall genomic divergence between presumptive taxa. Across many scaffolds, K2P distances averaged only 0.0041 ± 0.0080 for intraspecific comparisons between Desert and Prairie Massasaugas (Figure 3). As expected, interspecific comparisons between either Western Massasauga assembly and the Eastern Massasauga were larger (specifically, ~2X) at 0.0084 ± 0.0040. Finally, intergeneric K2P distances between *Crotalus* and *Sistrurus* species averged 0.0208 ± 0.0103, roughly 2.5X larger than the interspecific comparisons. Overall, these data are consistent with the idea that genomic divergence increases concomitantly with taxonomic divergence. All estimates for genetic distances between nuclear genome assemblies are depicted in Figure 3 and reported in Table 1. Alignment statistics are reported in Table A.3.
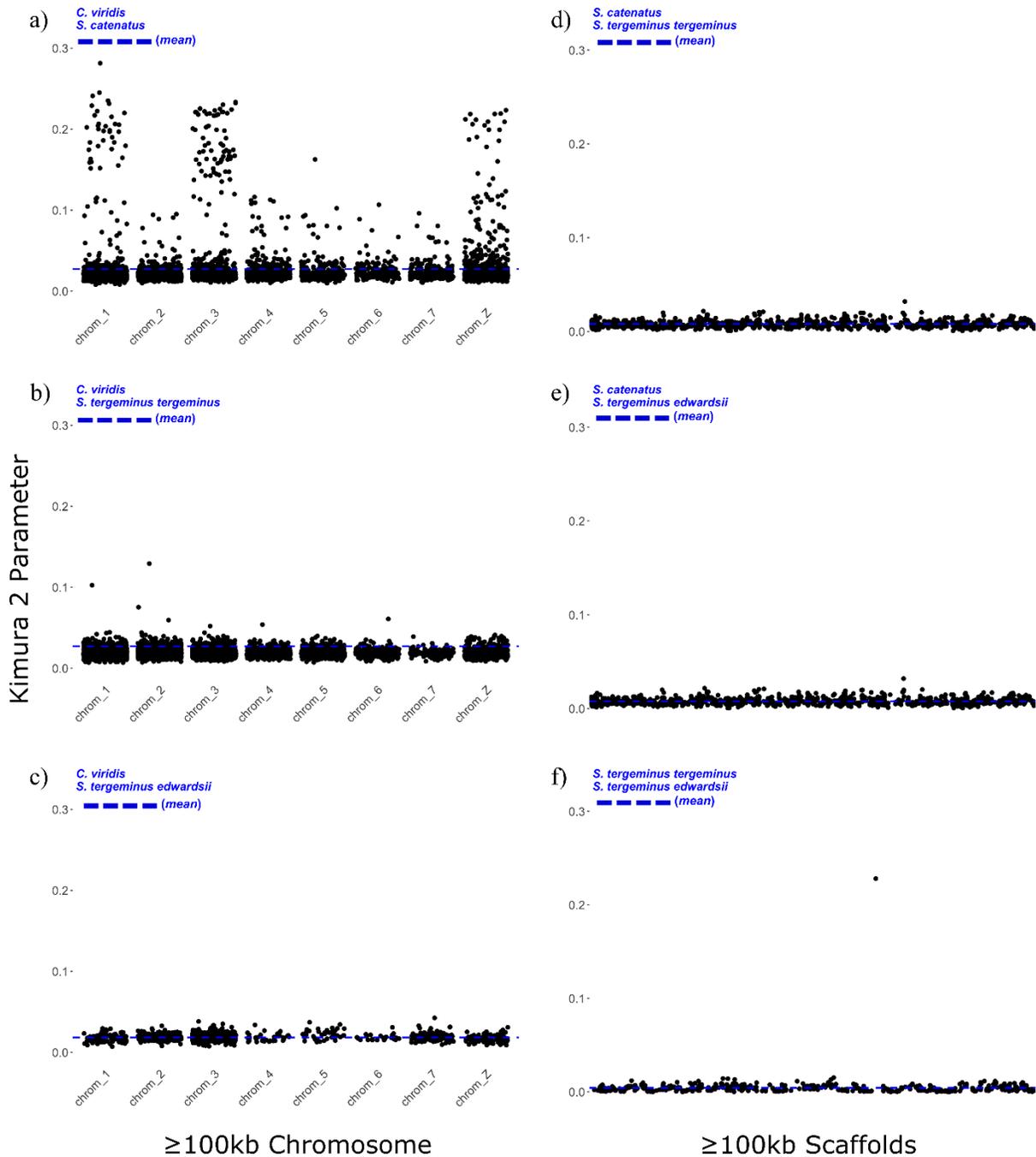
Figure 1.3. Distribution of pairwise genetic distance estimates (K2P; Kimura 1980) between ≥100kb scaffolds in nuclear genome assemblies: (a-c) represents K2P estimates derived from local alignments of ≥100kb scaffolds to the *C. viridis* chromosome-level assembly; d-e) represents K2P estimates from local alignments of ≥100kb *S. tergeminus* scaffolds to the *S catenatus* scaffold-level assembly; (f) K2P estimates derived from local alignments of ≥100kb scaffolds between putative *S. tergeminus* sub-species. Horizontal dashed lines in each panel identify the mean K2P value for each alignment

17

Table 1.1. Pairwise K2P genetic distances from rattlesnakes. NA, not available

Taxon 1	Taxon 2	Genetic Distance (K2P)	
		Mitochondrial	Nuclear
Sistrurus tergeminus tergeminus	*Sistrurus tergeminus edwardsii*	0.0103 ± 0.0008	0.00404 ± 0.01065
Sistrurus catenatus	*Sistrurus tergeminus edwardsii*	0.094 ± 0.0027	0.0078 ± 0.00331
Sistrurus catenatus	*Sistrurus tergeminus tergeminus*	0.0907 ± 0.0026	0.00802 ± 0.00318
Crotalus horridus	*Sistrurus tergeminus edwardsii*	0.1567 ± 0.0034	NA
Crotalus horridus	*Sistrurus tergeminus tergeminus*	0.1537 ± 0.0034	NA
Crotalus horridus	*Sistrurus catenatus*	0.1525 ± 0.0034	NA
Crotalus adamanteus	*Sistrurus tergeminus edwardsii*	0.1482 ± 0.0034	NA
Crotalus adamanteus	*Sistrurus tergeminus tergeminus*	0.1451 ± 0.0034	NA
Crotalus adamanteus	*Sistrurus catenatus*	0.1453 ± 0.0033	NA
Crotalus adamanteus	*Crotalus horridus*	0.1134 ± 0.0028	NA
Crotalus viridis	*Sistrurus tergeminus edwardsii*	NA	0.0182 ± 0.00449
Crotalus viridis	*Sistrurus tergeminus tergeminus*	NA	0.0189 ± 0.00488
Crotalus viridis	*Sistrurus catenatus*	NA	0.0238 ± 0.02784

The K2P distance for mitochondrial DNA between Desert and Prairie Massasaugas was 0.0103 ± 0.0008. Interspecific distances between Eastern and Western Massasauga mtDNA genomes were about nine-fold higher, 0.0940 ± 0.0027 and 0.0907 ± 0.0026 when comparing the Eastern Massasauga to the Desert and Prairie Massasauga sequences respectively. Intergenic distances between *Sistrurus* and *Crotalus* spp. ranged from 0.1451 ± 0.0034 (Prairie Massasauga – Eastern Diamondback) to 0.1567 ± 0.0034 (Desert Massasauga – Timber Rattlesnake).

1.3.3   SNP identification and genotyping

We tested 192 candidate SNPs and 184 SNPs were genotyped (i.e., only 8 SNPs failed to consistently amplify) in 78 samples (10 consistently failed to amplify). We subsequently removed 13 markers due to gametic phase disequilibrium. Our final set consisted of 171 markers, including 83 intergenic and 88 genic, 78 of which were from the Desert Massasauga and 93 from the Prairie Massasauga. As expected given our sampling regime, we found no unknown replicate samples and no first or second-degree relatives among our samples. Our genotyping error rate (e) was low, averaging 0.0082 (Desert: 0.0072, Prairie: 0.0091), and is heretofore ignored in population-level analyses.

Our genic markers include those within genes related to metabolism, venom production, and immune function among others (Table A.6). In contrast, our intergenic markers reside at least 10kb from known protein-coding genes. We partitioned our markers in this fashion to parse these two aspects of genomic variation if needed (e.g., due to signals of local adaptation). Similarly, half of the markers were developed from *S. t. tergeminus* reads and half from *S. t. edwardsii* reads to reduce taxon-specific ascertainment biases. Overall, however, our analyses revealed no substantive differences among the data partitions (i.e., ALL, HWE, GENIC, INTERGENIC, STT, and STE; Table 2) and thus for simplicity below we refer to results for ALL markers. For example, H_O was 0.300 ± 0.250 for all markers and did not differ between genic and intergenic markers (0.331 ± 0.245 and 0.267 ± 0.253; Fisher's exact test of 0.849, p>0.05). Deviations from HWE were no more common in the genic markers (Fisher's exact test of 0.480, p>0.05) than in the intergenic markers. We tested for the Wahlund effect for all subsets of data and found no such evidence given the mean F_{IS} for all markers was not significantly different than zero ($F_{IS} = 0.073 \pm 0.395$) and the slope of the relationship was greater than 1 (slope = 1.995 ± 0.583, adjusted $R^2 = 0.0659$).

Table 1.2. Observed and expected heterozygosity for various subsets of data. Columns display mean heterozygosity estimates with standard deviation for divisions of samples, including all individuals and then calculated independently based on subspecific designation. Overall, there are no obvious indications of ascertainment biases that would be expected if markers developed from one putative subspecies or population were deployed in a differentiated subspecies or population. See Methods and Figure B.1 for more details

Dataset	All Samples		Desert Massasaugas		Prairie Massasaugas	
	H_O	H_E	H_O	H_E	H_O	H_E
All Markers	0.300 ± 0.250	0.308 ± 0.173	0.291 ± 0.255	0.304 ± 0.179	0.311 ± 0.259	0.298 ± 0.177
Intergenic Markers	0.267 ± 0.253	0.283 ± 0.180	0.263 ± 0.262	0.279 ± 0.186	0.272 ± 0.254	0.276 ± 0.181
Genic Markers	0.331 ± 0.245	0.331 ± 0.165	0.317 ± 0.246	0.328 ± 0.169	0.347 ± 0.260	0.319 ± 0.171
Desert Markers	0.290 ± 0.216	0.320 ± 0.163	0.289 ± 0.215	0.324 ± 0.161	0.292 ± 0.237	0.296 ± 0.178
Prairie Markers	0.307 ± 0.277	0.298 ± 0.182	0.293 ± 0.285	0.288 ± 0.192	0.326 ± 0.277	0.300 ± 0.177

1.3.4 Population Genetic Analyses

We genotyped 78 individual snakes that were retained through filtering. Our Western Massasauga samples were collected across 7 US states (Arizona, Colorado, Kansas, Missouri, New Mexico, Oklahoma, Texas) and 1 Mexican state (Coahuila). For metadata of genotyped samples, see Table A.1. After Bonferroni correction, we found 32 out of 171 loci to be out of expected HWE proportions, 15 in heterozygote excess and 17 in a heterozygote deficit (Table A.6). For our population statistics, we found an overall mean F_{ST} across all markers of 0.0264 ± 0.0525 and 0.0308 ± 0.0548 for markers in HWE. Mean Fst between putative subspecies was 0.0318 ± 0.0612 for genic SNPs and 0.0198 ± 0.0414 for intergenic SNPs. F_{ST} also was calculated separately for SNPs selected from the reads of each subspecies (Desert mean = 0.0358 ± 0.0588; Prairie mean = 0.0178 ± 0.0458). Table 3 reports F_{ST} calculations for all data subsets.

Table 1.3. F_{ST} for all subsets of data. Rows indicate subsets of markers (see Methods and Figure B.1 for more details). Columns are as follows: dataset partition, mean F_{ST} between two populations defined by subspecific designation, then randomization assignments (500 trials) into two populations of each putative subspecies, Desert and Prairie. Means are shown with standard deviations. Overall, these data indicate that measurable genetic differentiation between putative subspecies is small and that, if the data are normally distributed, the 95% confidence intervals around mean F_{ST} values include zero

Dataset	Desert v. Prairie	Mean within Desert	Mean within Prairie
All Markers	0.0264 ± 0.0525	-0.000352 ± 0.00443	-0.00025 ± 0.00406
HWE Markers	0.0308 ± 0.0548	0.000126 ± 0.00537	0.000108 ± 0.00522
Intergenic Markers	0.0198 ± 0.0414	-0.00025 ± 0.00529	-0.00009 ± 0.00664
Genic Markers	0.0318 ± 0.0612	-0.00009 ± 0.00588	-0.00035 ± 0.00534
Desert Markers	0.0358 ± 0.0588	-0.00062 ± 0.00603	0.00016 ± 0.00663
Prairie Markers	0.0178 ± 0.0458	$0.00034\ 0.00531$	0.00018 ± 0.00543

We retained 16 of the axes from our PCA, explaining a cumulative 51% of the variation. The samples did not cluster according to subspecies status along any axis, though samples did show some separation according to geography on the primary two principal axes. Samples are plotted according to their position on the first two primary axes (11.3% of variation cumulatively) in Figure 4.

The Bayesian admixture analysis conducted in LEA determined optimum parameters $\alpha = 5$ and $k = 4$. There were no clear distinctions between subspecies (Figure 5). Additionally, we ran analyses for each subset of SNP data (genic, intergenic, desert, prairie) which had varying optimum alpha parameters, and output for all $k = 2$ through 5 for full data and subsets of data are shown in Figure B.2.

The Mantel test for autocorrelation of genetic and geographic distances was significant ($r = 0.181$, $p = 0.019$) for our subset of samples with geographic coordinates. These 37 individuals were collected from Kansas, Missouri, New Mexico, Oklahoma and Texas (Table A.1). Graphical results for the Mantel test are in Figure B.3.
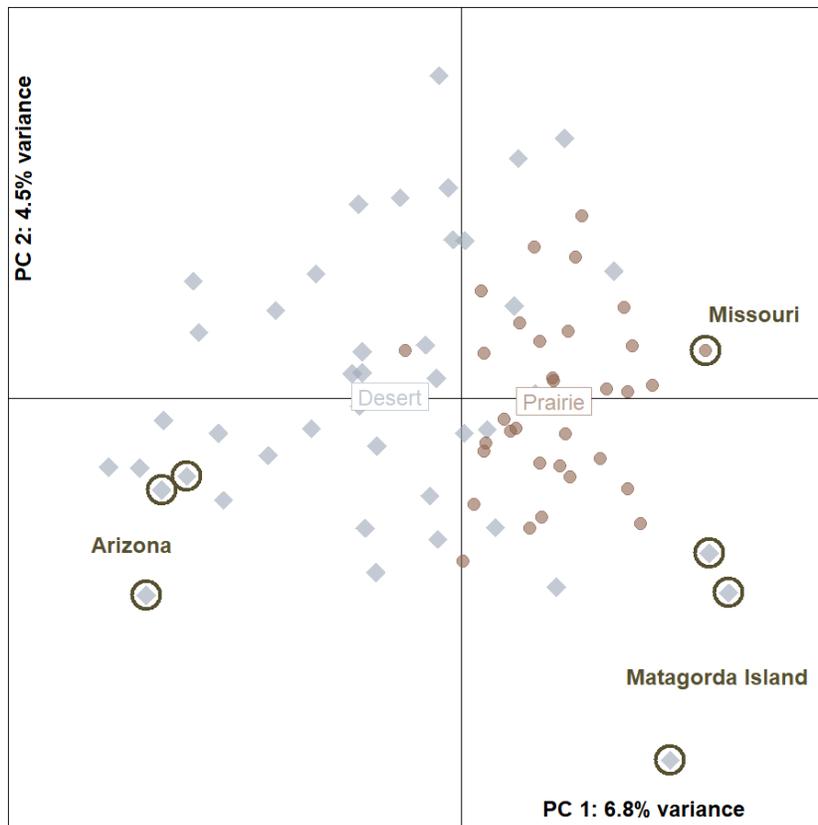
Figure 1.4. PCA of genotyped Western Massasauga samples. Desert (blue-grey diamonds) and Prairie (brown circles) samples are generally clustered in eigenspace. Cumulatively, PCs 1 and 2 account for 10.9% of the variance in the data. Consistent with a simple isolation-by-distance model, AZ and NM samples (Desert) fall further to the left of PC1 and CO (Desert) samples tend towards the top of PC2. All three fire mortalities from Matagorda Island (Desert) are distinctly in the lower right of the plot (circled), suggesting that these data have the capacity to identify isolated populations. As an example, we have circled the sole MO sample (Prairie) that is both farthest east and furthest on PC1 from the AZ (also circled) and NM samples of any Prairie Massasaugas. Overall, these data provide little if any support for subspecific designations in the Western Massasauga
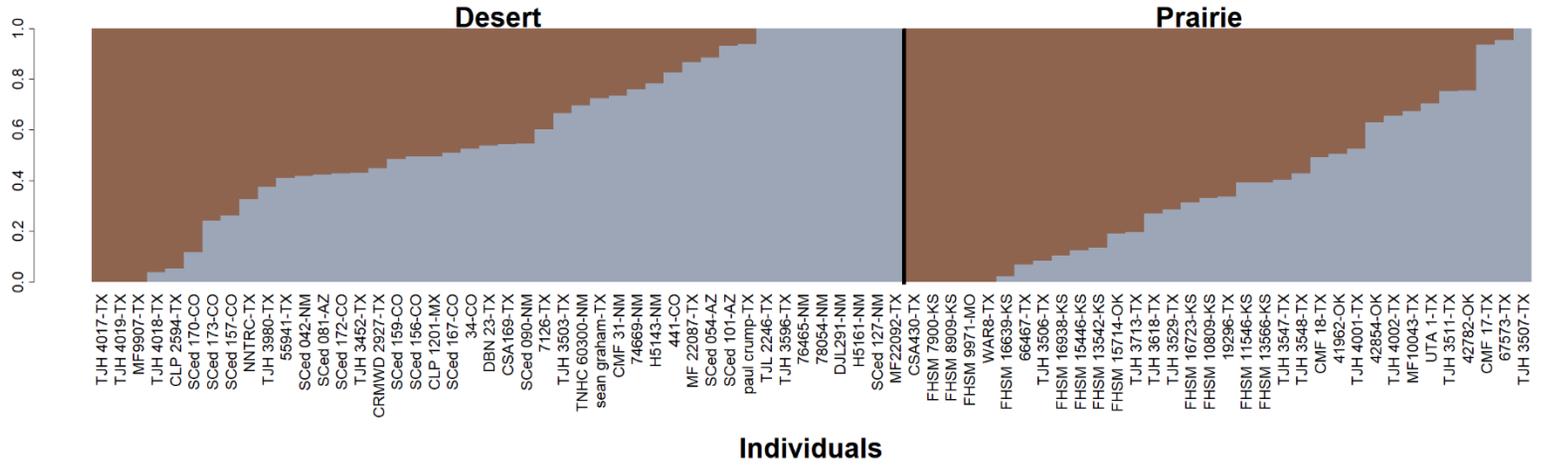
Figure 1.5. Bayesian admixture analysis for k = 2 from LEA using all SNP markers. Each column represents a single individual labelled with sample name and state, with nominal Desert and Prairie Massasaugas split left to right by the dark line, with samples for each putative subspecies sorted separately by admixture coefficient. The results shown here are most consistent with a single gene pool for Western Massasauga; if each nominal subspecies represented a distinct population segment, we would expect to see a genetic discontinuity that corresponded with the taxonomic designations. For other admixture plots, see Figure B.1

1.4 Discussion

Herein, we generated new discrete character (DNA sequences) and frequency-based marker datasets (SNP genotypes) to help provide conservation context to pending ESA decisions regarding the Desert Massasauga, *S. t. edwardsii*. The petition filed by WildEarth Guardians (2010) cites habitat degradation and loss (native habitat conversion, overgrazing, urbanization, desertification, water resource depletion, habitat fragmentation and isolation, and road-related mortality), increased death (intentional culls, vehicle-related deaths, and predation), and other factors (disease, naturally low survivability and fecundity, prey loss, drought and climate change, and pet trade collection) that contribute to population declines. Our data do not directly speak to many of these issues, but they do address whether the taxonomic entity *S. t. edwardsii* actually exists as a distinct population segment. If *S. t. edwardsii* and *S. t. tergeminus* are distinct enough that they merit different management strategies, then we expect those differences would manifest themselves by sundering the Western Massasauga gene pool. First, we might expect subspecific differences reflected in the genomes themselves (assembly sizes, TE content, GC content, etc.). Second, we would expect sequence differentiation (K2P) in both nuclear and mitochondrial genomes to be significantly greater than zero and consistent with values from pairs of other, well-established reptilian subspecies. Third, we would expect patterns of genetic differentiation (F_{ST}) to reveal obvious discontinuities between subspecies, including ascertainment biases where genetic differentiation was exaggerated when assessing one taxon with markers developed in the other taxon. Fourth, we would expect clusters of individuals from each subspecies to group distinctly in a PCA. Fifth, we would expect admixture analyses to reveal sharp departures from k=1 (i.e., strong support for k=2). Sixth, we would expect evidence of a Wahlund effect. Finally, we might see subspecific evidence of local adaptation. Below, we evaluate each of these lines of evidence in turn.

Snakes demonstrate a remarkable diversity in genome size and structure that challenge traditional notions of genome evolution (Pasquesi et al. 2018), and thus we might expect differences in basic genome statistics between two different subspecies. The assembly statistics in Table A.4 indicate quite similar genome compositions between both putative subspecies of Western Massasauga. The similarity in assembly size and GC content are indicators that Desert and Prairie Massasauga genomes are, at a gross level, very similar in structure. Clearly this is weak

evidence to suggest that the Western Massasauga is *not* composed of two distinct population segments, but the genomic similarities are certainly consistent with the stronger evidence below.

Kimura's 2-parameter distance between any two individuals should, in principle, be zero only for monozygotic twins or other clones (barring mutation). Thus, any two randomly chosen individuals from a population should have a K2P value that exceeds 0. Using nuclear DNA sequences, we estimated a K2P value of 0.0041 ± 0.0080 between individual *S. t. tergeminus* and *S. t. edwardsii*, a value near the expected lower bound of zero. Clearly, this value suggests very little genomic differentiation between our representatives of each putative subspecies. There is currently no "standard" level of genomic differentiation that corresponds to any level of taxonomic hierarchy, but it seems quite possible that one day this quantitative approach will be useful in the delineation of subspecies and higher taxa. Our survey of comparative K2P values in snake subspecies from the literature and publicly available nuclear gene sequences averaged 0.004 ± 0.005, quite similar to our estimates in Western Massasauga. We note that our literature-based K2P estimates are biased by the small number of genes considered (i.e., most studies were not "genomic"), so further research will be required to firmly document the degree of genomic differentiation between putative subspecies. In the meantime, the genomic K2P evidence supports the idea of exceedingly low differentiation among Western Massasaugas and provides limited or no support for subspecific delineation.

Snake mitochondrial sequence data are far more publicly available than snake nuclear sequence data, and thus more taxonomic context is available. Our mitochondrial K2P estimate between putative Desert and Prairie Massasauga was 0.0103 ± 0.0008 (Figure 3; Table 1), far below estimates of subspecies-level distances in *Macroprotodon* spp. (0.093 ± 0.049) (Carranza et al. 2004) and *Psammophis* spp. (0.067 ± 0.033) (Kelly et al. 2008). Instead, the estimate between the Desert and Prairie mitochondrion corresponds to the range of mean inter-population mtDNA distances for different populations of *Psammophis* (0.016 ± 0.016; Kelly et al. 2008) and between populations of *Natrix maura* (0.042 ± 0.008; Guicking et al. 2008). Overall, our sequence data indicate that the mtDNA divergence between putative subspecies of Western Massasauga is far less than what might be expected of distinct subspecies but falls well within the range of population-level divergence. In other words, our mtDNA data are generally consistent with the idea of a single Western Massasauga taxon that contains modest levels of nucleotide variability but generally inconsistent with data from comparisons between established snake subspecies.

Our overall F_{ST} value for all nuclear SNP markers between the two subspecies was low ($F_{ST} = 0.0264 \pm 0.0525$), but similar to an independent subspecific comparison of snake taxa ($F_{ST} = 0.02 – 0.08$ for *Micrurus diastema* ssp.; Reyes-Velasco et al. 2020). However, we note that a) our geographic sampling was extensive (e.g., our two most distant samples were separated by 1,300 km); b) the two putative subspecies have mostly disparate geographic ranges; and c) the little differentiation we observed in F_{ST} values better reflects IBD (see below) than subspecific delineation. Irregardless, it is clear from these F_{ST} results that a very small proportion of the genetic variation within Western Massasaugas differentiates putative subspecies.

Our PCA results reveal little differentiation between putative subspecies (Figure 4). If there were genetic differentiation between the two subspecies in question, we would see samples from each subspecies clustering separately on at least one axis. However, Desert and Prairie Massasaugas were largely coincident along all axes. Samples ordinated slightly with geography along axes 1 and 2, with those from populations at the range limits (i.e. AZ, NM, CO, MO, Matagorda Island) tending to ordinate further along axes 1 and 2 with sympatric samples. An exception to this is the lone sample from Mexico, which did not separate from the main grouping of samples.

Given that samples mostly sort along the first two primary PCA axes by geography, we conducted a Mantel test of autocorrelation between genetic and geographic distances. The significance of the Mantel test indicates that the gene pool of Western Massasaugas is shaped by patterns of IBD. Overall, we think the most parsimonious interpretation of the genetic structure results is simple IBD, and that the PCA data do not provide convincing support for the genetic distinctiveness of formal taxonomic subspecies. This should not be too surprising given the low motility and dispersal capacity of Massasaugas, the large geographic range sampled herein, and the fact that IBD is the *de facto* null hypothesis in population genetics.

Next, we consider the results of our admixture analyses. LEA results reveal no evidence of strong population structure within Western Massasaugas and, instead, are more consistent with genetic homogeneity, albeit with some IBD, across the sampled range. If the two putative subspecies were genetically differentiated, one would expect that genetic assignment tests could reliable identify a Western Massasauga of unknown origin to one or the other subspecies but Figure 5 illustrates how assignment probabilities would be virtually identical. Additionally, if each nominal subspecies represented a distinct population segment, we would expect to see a genetic

discontinuity that corresponded with the taxonomic designations in Figure 5, but that is not the case. Evaluation of admixture scenarios where k > 2 would require much more intensive sampling across the range, but we think such scenarios are unlikely because if two or more genetically structured populations are artificially grouped, even if the subpopulations are in HWE, we would expect a Wahlund effect. For example, if the two subspecies were genetically differentiated but we analyzed them as a single unit, we should find an overall deficit of heterozygotes, but we found no such evidence in our analyses (Table A.7). This is an admittedly weak test for "distinct populations", as it is in effect an absence of evidence, but it is consistent with our interpretation of our other analyses.

Finally, we see little evidence for subspecific designation based on the possibility of local adaptation because there is no evidence of strong differential selection on our genic markers; they are no more likely to deviate from HWE than the putatively neutral intergenic markers. We again explicitly acknowledge this is a weak test of population distinction, but there is no obvious taxonomic signal of local adaptation in our dataset.

These data represent the most complete genetic survey of Western Massasaugas available, but our study of course has limitations. We have no ecological (e.g., diet), behavioral (e.g., mate choice), or physiological (e.g., thermal tolerance) data. Furthermore, our genetic data are not ideal as most of our samples were collected via driving surveys and found dead on the road, therefore yielding fragmented DNA unsuitable for many assays (e.g., RadSeq). The Fluidigm SNPtype assay works remarkable well with poor quality DNA (Carroll et al. 2018; von Thaden et al. 2020), but in the end it surveys relatively few markers across the genome. Despite significant effort, we were likewise limited in the number of biological samples we surveyed across a very wide geographic range. All of these limitations add noise to our dataset, but the overall consistency among our various analyses (e.g., genetic/genomic, and frequency/categorical) speak to the significant biological signal that nevertheless remains.

1.5 Conclusions

Our data on the Western Massasauga are generally uniform across different data types and different analyses, revealing no obvious genetic discontinuities that yield a distinct population segment. These genetic and genomic data do not support the idea of *either* Desert or Prairie Massasauga; instead, our data suggest that Western Massasaugas consist of a single, relatively

diverse gene pool. Previous evidence sharply distinguished the Western Massasauga from its Eastern sister species (Kubatko et al. 2011; Ryberg et al. 2015). Those studies cast initial doubt on the traditional division of Desert and Prairie Massasaugas as separate subspecies but were lacking power due to restricted sampling. This study, with greater geographic, genetic, and genomic resolution, buttresses previous studies of population structure in Massasaugas (Kubatko et al. 2011; Ryberg et al. 2015). Therefore, based on the genetic and genomic data previously published and herein, we recommend that Western Massasaugas, *Sistrurus tergeminus*, not be considered as two genetically differentiated subspecies but as a genetically unified species. Our data also illustrate the ability of genetics and genomics to help delineate taxa, in this case unifying artificial taxonomic constructs that do not reflect biological realities.

1.6   Data Availability

Sequencing reads and assemblies have been submitted to NCBI under BioProject accession PRJNA668351. Analyses were conducted in R or unix environments and scripts are available in the online resources of the publication of the same name.

1.7   References

Anderson CD, Gibbs HL, Douglas ME, Holycross AT (2009) Conservation genetics of the desert massasauga rattlesnake (*Sistrurus catenatus edwardsii*). Copeia 2009:740–747. https://doi.org/10.1643/cg-08-152

Andrews S (2010) FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Blair C, Sánchez-Ramírez S (2016) Diversity-dependent cladogenesis throughout western Mexico: evolutionary biogeography of rattlesnakes (Viperidae: Crotalinae: *Crotalus* and *Sistrurus*). Mol Phylogenet Evol 97: 145–154. https://doi.org/10.1016/j.ympev.2015.12.020

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

Broad Institute (2018) Picard Tools version 2.18.2. Broad Institute. GitHub repository http://broadinstitute.github.io/picard/.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. BMC Bioinformatics 10:1–9. https://doi.org/10.1186/1471-2105-10-421

Carranza S, Arnold EN, Wade E, Fahd S (2004) Phylogeography of the false smooth snakes, *Macroprotodon* (Serpentes, Colubridae): mitochondrial DNA sequences show European populations arrived recently from northwest Africa. Mol Phylogenet Evol 33:523–532. https://doi.org/10.1016/j.ympev.2004.07.009

Carroll EL, Bruford MW, DeWoody JA, Leroy G, Strand A, Waits L, Wang J (2018) Genetic and genomic monitoring with minimally invasive sampling methods. Evol Appl 11:1094–1119. https://doi.org/10.1111/eva.12600

Chikhi R, Medvedev P ()2014 Informed and automated *k*-mer size selection for genome assembly. *Bioinformatics* 30(1):31–37. https://doi.org/10.1093/bioinformatics/btt310

Clark RW, Brown WS, Stechert R, Zamudio KR (2010) Roads, interrupted dispersal, and genetic diversity in timber rattlesnakes. Conserv Biol 24:1059–1069. https://doi.org/10.1111/j.1523-1739.2009.01439.x

Clark RW, Marchand MN, Clifford BJ, Stechert R, Stephens S (2011) Decline of an isolated timber rattlesnake (*Crotalus horridus*) population: interactions between climate change, disease, and loss of genetic diversity. Biol Conserv 144:886–891. https://doi.org/10.1016/j.biocon.2010.12.001

Colley M, Lougheed SC, Otterbein K, Litzgus JD (2017) Mitigation reduces road mortality of a threatened rattlesnake. Wildlife Res 44:48–59. https://doi.org/10.1071/WR16130

Crother BI, Savage JM, Holycross AT (2011) Case 3571 *Crotalinus catenatus* Rafinesque, 1818 (currently *Sistrurus catenatus*) and *Crotalus tergeminus* Say in James, 1822 (currently *Sistrurus tergeminus*; Reptilia, Serpentes): proposed conservation of usage by designation of neotypes for both species. Bull Zool Nomencl 68:271–274. https://doi.org/10.21805/bzn.v68i4.a3

Crother BI, Savage JM, Holycross AT (2013) *Crotalinus catenatus* Rafinesque, 1818 (currently *Sistrurus catenatus*) and *Crotalus tergeminus* Say in James, 1822 (currently *Sistrurus tergeminus*; Reptilia, Serpentes): usage conserved by designation of neotypes for both species. Bull Zool Nomencl 70:271–274. https://doi.org/10.21805/bzn.v68i4.a3

Doyle JM, Katzner TE, Roemer GW, Cain JW III, Millsap BA, McIntyre CL, Sonsthagen SA, Fernandez NB, Wheeler M, Bulut Z, Bloom PH, DeWoody JA (2016) Genetic structure and viability selection in the golden eagle (*Aquila chrysaetos*), a vagile raptor with a Holarctic distribution. Conserv Genet 17:1307–1322. https://doi.org/10.1007/s10592-016-0863-0

Fernandes F, Pereira L, Freitas AT (2009) CSA: an efficient algorithm to improve circular DNA multiple alignment. BMC Bioinformatics 10**:**230. https://doi.org/10.1186/1471-2105-10-230

Fetzner JW Jr (1999) Extracting high-quality DNA from shed reptile skins: a simplified method. Biotechniques 26:1052–1054. https://doi.org/10.2144/99266bm09

Fitzgerald LA, Painter CW (2000) Rattlesnake commercialization: long-term trends, issues, and implications for conservation. Wildlife Soc B 28:235–253.

Frichot E, François O (2015) LEA: an R package for landscape and ecological association studies. Methods Ecol Evol 6(8):925–929. https://doi.org/10.1111/2041-210X.12382

Gibbs HL, Murphy M, Chiucchi JE (2011) Genetic identity of endangered massasauga rattlesnakes (*Sistrurus* sp.) in Missouri. Conserv Genet 12:433–439. https://doi.org/10.1007/s10592-010-0151-3

Gloyd H (1955) A review of the massasauga, *Sistrurus catenatus*, of the southwestern United States (Serpentes: Crotalidae). Bull Chi Acad Sci 10:83–98.

Gregory TR (2020) Animal Genome Size Database. Available online at: http://www.genomesize.com

Guicking D, Joger U, Wink M (2008) Molecular phylogeography of the viperine snake *Natrix maura* (Serpentes: Colubridae): evidence for strong intraspecific differentiation. Org Divers Evol 8:130–145. https://doi.org/10.1016/j.ode.2007.05.001

Hahn C, Bachmann L, Chevreux B (2013) Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. Nucleic Acids Research 41(13):129, https://doi.org/10.1093/nar/gkt371

Holycross AT, Anton TG, Douglas ME, Frost DR (2008) The type localities of *Sistrurus catenatus* and *Crotalus viridis* (Serpentes: Viperidae), with the unraveling of a most unfortunate tangle of names. Copeia 2008:421–424. https://doi.org/10.1643/ch-07-095

Holycross AT, Mackessy SP (2002) Variation in the diet of *Sistrurus catenatus* (massasauga), with emphasis on *Sistrurus catenatus edwardsii* (desert massasauga). J Herpetol 36:454–464. https://doi.org/10.1670/0022-1511(2002)036[0454:VITDOS]2.0.CO;2

Jackman SD, Vandervalk BP, Mohamadi H, Chu J, Yeo S, Hammond SA, Jahesh G, Hamza Khan, Coombe L, Warren RL, and Birol I (2017). ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter**.** *Genome research*, 27(5):768-777. doi:10.1101/gr.214346.116

Jombart T (2008) Adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24:1403–1405. https://doi.org/10.1093/bioinformatics/btn129

Kelly CMR, Barker NP, Villet MH, Broadley DG, Branch WR (2008) The snake family Psammophiidae (Reptilia: Serpentes): phylogenetics and species delimitation in the African sand snakes (*Psammophis* Boie, 1825) and allied genera. Mol Phylogenet Evol 47:1045–1060. https://doi.org/10.1016/j.ympev.2008.03.025

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120. https://doi.org/10.1007/BF01731581

Knight A, Styer D, Campbell JA, Pelikan S, Densmore LD, Mindell DP (1993) Choosing among hypotheses of rattlesnake phylogeny a best fit rate test for DNA sequence data. Syst Biol 42:356–367. https://doi.org/10.1093/sysbio/42.3.356

Kubatko LS, Gibbs HL, Bloomquist EW (2011) Inferring species-level phylogenies and taxonomic distinctiveness using multilocus data in *Sistrurus* rattlesnakes. Syst Biol 60:393–409. https://doi.org/10.1093/sysbio/syr011

Kumar S, Stecher G, Li M, Knyaz C, Tamura K (2018) MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol 35:1547–1549. https://doi.org/10.1093/molbev/msy096

Kumar S, Stecher G, Suleski M, Hedges SB (2017) TimeTree: a resource for timelines, timetrees, and divergence times. Mol Biol Evol 34:1812–1819. https://doi.org/10.1093/molbev/msx116

Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv. http://arxiv.org/abs/1303.3997

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu SM, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam TW, Wang J (2012) SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *GigaScience* 1(1):2047–217X–1–18, https://doi.org/10.1186/2047-217X-1-18

Mackessy SP, Keinath DA, Mcgee M, Mcdonald D, Ise T (2005) Desert massasauga rattlesnake (*Sistrurus catenatus edwardsii*): a technical conservation assessment. https://www.fs.usda.gov/Internet/FSE_DOCUMENTS/stelprdb5182073.pdf. Accessed 22 September 2020.

McCluskey EM, Bender D (2015) Genetic structure of western massasauga rattlesnakes (*Sistrurus catenatus tergeminus*) . J Herpetol 49:343–348. https://doi.org/10.1670/14-016

Murphy RW, Fu J, Lathrop A, Feltham JV, Kovac V (2002) Phylogeny of the rattlesnakes (*Crotalus* and *Sistrurus*) inferred from sequences of five mitochondrial DNA genes. In: Schuett GW, Hoggren M, Douglas ME, Greene HW (eds) Biology of the vipers. Eagle Mountain Publishing LC, Utah, pp 66–92

O'Brien SJ, Mayr E (1991) Bureaucratic mischief: recognizing endangered species and subspecies. Science 251:1187–1188. https://doi.org/10.1126/science.251.4998.1187

Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290. https://doi.org/10.1093/bioinformatics/btg412

Pasquesi GIM, Adams RH, Card DC, Schield DR, Corbin AB, Perry BW, Reyes-Velasco J, Ruggiero RP, Vandewege MW, Shortt JA, Castoe TA (2018) Squamate reptiles challenge paradigms of genomic repeat element evolution set by birds and mammals. Nat Commun 9:2774. https://doi.org/10.1038/s41467-018-05279-1

Reyes-Velasco J, Adams RH, Boissinot S, Parkinson CL, Campbell JA, Castoe TA, Smith EN (2020) Genome-wide SNPs clarify lineage diversity confused by coloration in coralsnakes of the *Micrurus diastema* species complex (Serpentes: Elapidae). Mol Phylogenet Evol 147:106770. https://doi.org/10.1016/j.ympev.2020.106770

Rousset F (2008) GENEPOP'007: A complete re-implementation of the GENEPOP software for Windows and Linux. Mol Ecol Resourc 8:103–106. https://doi.org/10.1111/j.1471-8286.2007.01931.x

Ryberg WA, Harvey JA, Blick A, Hibbitts TJ, Voelker G (2015) Genetic structure is inconsistent with subspecies designations in the western massasauga *Sistrurus tergeminus*. J Fish Wildl Manag 6:350–359. https://doi.org/10.3996/122014-JFWM-093

Schield DR, Card DC, Hales NR, Perry BW, Pasquesi GM, Blackmon H, Adams RH, Corbin AB, Smith CF, Ramesh B, Demuth JP, Bertrán E, Tollis M, Meik JM, Mackessy SP, Castoe TA (2019) The origins and evolution of chromosomes, dosage compensation, and mechanisms underlying venom regulation in snakes. Genome Res 29:590–601. https://doi.org/10.1101/gr.240952.118

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673-4680.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S, DePristo MA (2013) From fastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Curr Protoc Bioinform 43:11.10.1–11.10.33. https://doi.org/10.1002/0471250953.bi1110s43

von Thaden A, Nowak C, Tiesmeyer A, Reiners TE, Alves PC, Lyons LA, Mattucci F, Randi E, Cragnolini M, Galián J, Hegylei Z, Kitchener AC, Lambinet C, Lucas JM, Mölich T, Ramos L, Schockert V, Cocchiararo B (2020) Applying genomic data in wildlife monitoring: development guidelines for genotyping degraded samples with reduced single nucleotide polymorphism (SNP) panels. Mol Ecol Resourc 20:662–680. https://doi.org/10.1111/1755-0998.13136

Waples RS, Allendorf F (2015) Testing for Hardy-Weinberg proportions: have we lost the plot? J Hered 106:1–19. https://doi.org/10.1093/jhered/esu062

Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York

Wild Earth Guardians (2010) Petition to list the desert massasauga (*Sistrurus catenatus edwardsii*) under the U.S. Endangered Species Act. https://pdf.wildearthguardians.org/site/DocServer/listing_petition_desert_massasauga.pdf?docID=685&AddInterest=1059. Accessed 22 September 2020.

APPENDIX A. ADDITIONAL TABLES

Table A.1. Genotyped and sequenced Sistrurus samples. Samples are sorted by species, then putative subspecies, and finally by county of collection. Individuals that did not successfully genotype at greater than 80% of loci are not included. The two representative individuals chosen for whole genome sequencing of Desert and Prairie Massasaugas and subsequent variant calling are labelled

(see supplemental file)

Table A.2 Sequence data and associated details gathered from publicly available, online resources

Taxa	Sequence type	Use in methods	Accession number	Citation
Crotalus adamanteus	mitochondrial genome	genetic distance estimates	NC_041524.1	Wu 2019
Crotalus horridus	mitochondrial genome	genetic distance estimates	NC_014400.1	Hall et al. 2013
Crotalus horridus	nuclear genome	genetic distance estimates	GCA_001625485.1	Rhoads et al. 2016
Crotalus pyrrhus	nuclear genome	genetic distance estimates	GCA_000737285.1	Reed College 2014
Crotalus viridis	nuclear genome	SNP selection & genetic distance estimates	GCA_003400415.2	Schield et al. 2019
Emydocephalus ijimae	nuclear genome	genetic distance estimates	GCA_004319985.1	Kishida et al. 2019
Hydrophis cyanocinctus	nuclear genome	genetic distance estimates	GCA_004023725.1	Second Military Medical University 2018
Hydrophis hardwickii	nuclear genome	genetic distance estimates	GCA_004023765.1	Second Military Medical University 2018
Hydrophis melanocephalus	nuclear genome	genetic distance estimates	GCA_004320005.1	Kishida et al. 2019

Laticauda colubrina	nuclear genome	genetic distance estimates	GCA_004320045.1	Kishida et al. 2019
Laticauda laticaudata	nuclear genome	genetic distance estimates	GCA_004320025.1	Kishida et al. 2019
Notechis scutatus	nuclear genome	genetic distance estimates	GCA_900518725.1	University of New South Wales 2018b
Ophiophagus hannah	nuclear genome	genetic distance estimates	GCA_000516915.1	Vonk et al. 2013
Pantherophis guttatus	nuclear genome	genetic distance estimates	GCA_001185365.2	Ullate-agote et al. 2014
Pantherophis obsoletus	nuclear genome	genetic distance estimates	GCA_012654085.1	University of Geneva 2020
Protobothrops flavoviridis	nuclear genome	genetic distance estimates	GCA_003402635.1	Shibata et al. 2018
Protobothrops mucrosquamatus	nuclear genome	genetic distance estimates	GCA_001527695.3	Aird et al. 2017
Pseudonaja textilis	nuclear genome	genetic distance estimates	GCA_900518735.1	University of New South Wales 2018a
Sistrurus miliarius	nuclear DNA sequences	genetic distance estimates	FJ659860-FJ660411	Gibbs and Diaz 2010
Sistrurus t. edwardsii	ND2 mitochondrial gene sequence	mitochondrial genome assembly	GQ359813.1	King and Ray 2009
Sistrurus t. tergeminus	ND2 mitochondrial gene sequence	mitochondrial genome assembly	GQ359814.1	King and Ray 2009
Thamnophis elegans	nuclear genome	genetic distance estimates	GCA_009769535.1	Bronikowski et al. 2019
Thamnophis sirtalis	nuclear genome	genetic distance estimates	GCA_001077635.2	WUGSC 2015

Table A.3 Alignment statistics for K2P calculations from Sistrurus and Crotalus samples.

(see supplemental file)

Table A.4 Sequencing statistics. Statistics for both representative individuals sequenced across the whole genome. Estimated genome size from kmergenie and used to calculate coverage.

	Desert Massasauga, *S.t. edwardsii*		**Prairie Massasauga**, *S.t. tergeminus*	
Estimated Genome Size	1,897,989,586 bp		1,937,494,987 bp	
	PE Reads	**MP Reads**	**PE Reads**	**MP Reads**
Total reads sequenced	481,905,794	136,670,652	488,785,880	206,014,690
Total bases sequenced	72,615,142,853	16,607,421,992	73,662,175,340	24,212,091,803
Estimated coverage	38.25x	8.75x	38x	12.5x
Average read length	150 bp	150 bp	150 bp	150 bp
Average GC content	39.5%	40.5%	39.5%	40%

Table A.5 Assemby N50 statistics. Results for assemblies attempted for Desert and Prairie Massasauga subspecies. N50 values for all assembly attempts. N50 statistics were calculated first with all scaffolds included and then with a minimum scaffold length of 200bp. The N50 value represents the median scaffold length; half of all scaffolds are longer than the N50 and half are shorter. As such, it represents one metric of genome assembly quality.

	Desert Massasauga, *S.t. edwardsii*		**Prairie Massasauga**, *S.t. tergeminus*	
ABySS	N50	N50 (scaffolds > 200 bp)	N50	N50 (scaffolds > 200 bp)
k=30	576	1893	555	2266
k=50	1583	2048	3022	7412
k=60	3232	6160	10364	22979
k=70	2953	4238	6754	11941
k=80	4668	6099	7636	10837
k=100	4249	4771	6253	7284
SOAPdenovo2				
k=60	4616	7747	7563	13137

Table A.6 Genotyped SNPs that passed filtering protocols. Source sample and type of marker are shown (i.e. putatively neutral/intergenic, putatively non-neutral/adaptive/genic). Ontology and BLAST results are based on gene annotations and sequence synteny in the BLAST database. The last column indicates whether a locus is in or out of HWE.

(see supplemental file)

Table A.7 Wahlund test results for each data partition. Slope and adjusted R2 are for the relationship of FIS/FST.

marker subset	mean Fst	mean slope	adjusted R^2
all	0.073 ± 0.395	1.995 ± 0.583	0.0659
intergenic	0.115 ± 0.421	2.614 ± 1.139	0.0560
genic	0.035 ± 0.367	1.847 ± 0.647	0.0831
desert	0.127 ± 0.348	1.503 ± 0.681	0.0524
prairie	0.027 ± 0.428	2.477 ± 0.994	0.0604
intergenic desert	0.239 ± 0.282	1.447 ± 0.972	0.0355
genic desert	0.023 ± 0.374	1.832 ± 0.884	0.0838
intergenic prairie	0.007 ± 0.491	3.771 ± 2.248	0.0456
genic prairie	0.044 ± 0.366	1.991 ± 0.991	0.0674

References

Aird SD, Arora J, Barua A, Qiu L, Terada K, Mikheyev AS (2017) GBE Population Genomic Analysis of a Pitviper Reveals Microevolutionary Forces Underlying Venom Chemistry 9(10):2640–2649. https://doi.org/10.1093/gbe/evx199

Bronikowski A, G10K, Rockefeller Vertebrate Genome Lab, Dovetail Genomics (2019) *Thamnophis elegans* (Western terrestrial garter snake) genome, rThaEle1, primary haplotype.

Gibbs HL, Diaz J (2010) Identification of single copy nuclear DNA markers for North American pit vipers. Mol Ecol Resour10(1):177-180. doi: 10.1111/j.1755-0998.2009.02707.x. Epub 2009 Apr 23. PMID: 21565006.

Hall JB, Cobb VA, Cahoon AB (2013) The complete mitochondrial DNA sequence of *Crotalus horridus* (timber rattlesnake) Mitochondrial DNA 24 (2), 94-96 (2013)

King RB and Ray JW (2017) Direct Submission: Submitted (07-JUL-2009) Biological Sciences, Northern Illinois University, Montgomery Hall, Dekalb, IL 60115, USA

Kishida T, Go Y, Tatsumoto S, Tatsumi K, Kuraku S, Toda M (2019) Loss of olfaction in sea snakes provides new perspectives on the aquatic adaptation of amniotes. Proceedings of the Royal Society B: Biological Sciences 286(1910). https://doi.org/10.1098/rspb.2019.1828

Reed College. (2014) Draft assembly of whole genome sequence data for *Crotalus mitchelli pyrrhus* (the speckled rattlesnake), a member of the *Crotalus mitchellii* species complex (Viperidae) distributed throughout southwestern North America.

Rhoads Dand University of Arkansas (2016) *Crotalus horridus* genome sequence assembly and annotation. Collaborative project of the MCBIOS.

Schield DR, Card DC, Hales NM, Perry BW, Pasquesi GIM, Blackmon H, Adams RH, Corbin AB, Smith CF, Ramesh B, Demuth JP, Betran E, Tollis M, Meik JM, Mackessy SP, and Castoe TA (2019) The origins and evolution of chromosomes, dosage compensation, and mechanisms underlying venom regulation in snakes. Genome Research 29: 590-601.

Second Military Medical University (2018) De novo whole genome sequencing of *Hydrophis cyanocinctus* and *Hydrophis hardwickii*, two sea snake species from the South China Sea.

Shibata H, Chijiwa T, Oda-ueda N, Nakamu H (2018) The habu genome reveals accelerated evolution of venom protein genes. 1–11. https://doi.org/10.1038/s41598-018-28749-4

The Genome Institute at Washington University School of Medicine (WUGSC) (2015) *Thamnophis sirtalis* Genome sequencing and assembly.

Ullate-agote A, Milinkovitch MC, Tzika AC (2015) The genome sequence of the corn snake (*Pantherophis guttatus*), a valuable resource for EvoDevo studies in squamates. 888:881–888. https://doi.org/10.1387/ijdb.150060at

University of Geneva (2020) Genome assembly of the rat snake *Pantherophis obsoletus*.

University of New South Wales (2018a) De novo whole genome sequencing of the eastern brown snake, *Pseudonaja textilis*.

University of New South Wales (2018b) De novo whole genome sequencing of the mainland tiger snake, *Notechis scutatus*.

Vonk FJ, Casewell NR, Henkel CV, Heimberg AM, Jansen HJ, Woods AE, … De Koning APJ (2013) The king cobra genome reveals dynamic gene evolution and adaptation in the snake venom system. 110(51):20651–20656. https://doi.org/10.1073/pnas.1314702110

Wu S-X (2019) Characterization of the complete mitochondrial genome of *Crotalus adamanteus* (Eastern diamondback rattlesnake) Mitochondrial DNA B Resour 4(1):632-634

APPENDIX B. ADDITIONAL FIGURES
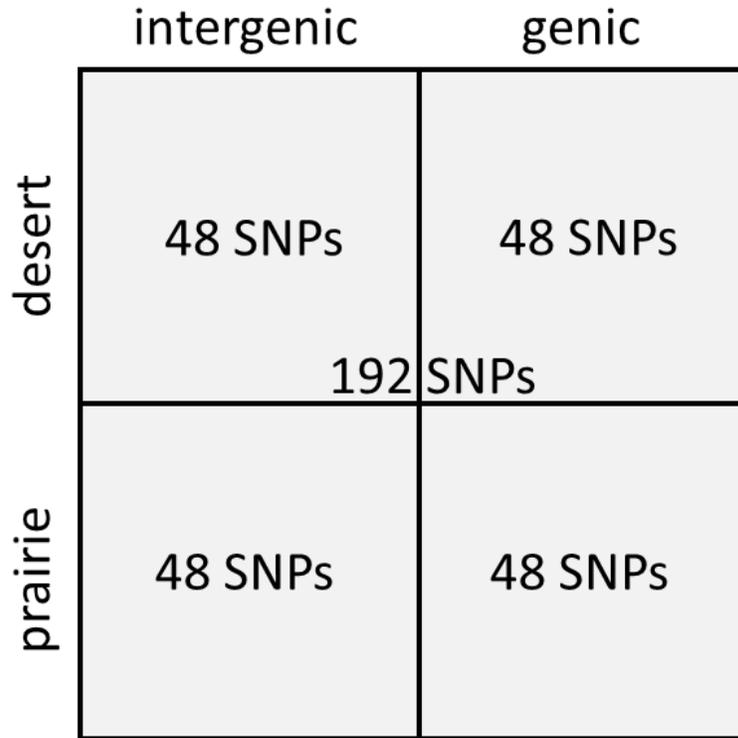
Fig. B.1 Diagram of the different categories of SNPs used for genotyping. SNPs are split in half by subspecies designation based on the source of reads used for SNP discovery and validation (i.e. Desert and Prairie). These subsets are again divided in half based on the region of the genome that the SNP was detected in. SNPs from within annotated exons of putative genes are considered potentially "adaptive" or "non-neutral" loci. SNPs from intergenic regions are considered "neutral" loci based on their distance (>10kb) from annotated protein-coding genes

(see supplemental file)

Fig. B.2 Admixture plots for all subsets of data based on marker type. Runs were conducted for optimum values of the alpha parameter at optimum k and surrounding values of k. Note that Desert and Prairie Massasaugas show little-to-no differentiation in all runs

(see supplemental file)

Fig. B.3 Results of the Mantel test for isolation by distance. Shown are a) a scatter plot of genetic distances plotted by geographic distances and b) a null distribution generated from randomized trials of all distances with the actual relationship of genetic-to-geographic distance marked with a diamond. Results of the Mantel test are indicative of a pattern of isolation-by-distance as expected for a low motility species with such a wide natural range

APPENDIX C. SCRIPTS

fastqc.sh - code to check wgs read quality

kmergenie.jellyfish.sh - code to estimate genome size and optimum kmer with kmergenie and jellyfish programs

abyss.denovo.sh - code for genome assembly of Sistrurus tergeminus tergeminus (Prairie Massasauga)

SOAPdenovo.sh - cod for genome assembly of Sistrurus tergeminus edwardsii (Desert Massasauga)

blast.sh - code to search in genome assembly for mitochondrial scaffold with mitochondrial sequence

check_read_alignments.sh - code to align wgs reads to reference mitochondrion and mitochondrial assembly to visually check data

mitobim.sh - code to generate mitochondrial assembly from wgs reads and a reference mitochondrion sequence

template_blast.sh - BLAST code for aligning genomes of two different individuals. Individualized and repeated serially for all pairwise comparisons.

calc_div.sh - bash submission file for calc_div.R to calculate K2P distances between full nuclear genomes

calc_div.R - R code for calculating full nuclear genome divergence (K2P distances) from BLAST results (template_blast.sh)

k2p_figure.sh - bash and R code for generation of genetic divergence figure (Figure 3) from full nuclear genome alignments and genetic distance calculations (template_blast.sh; calc_div.sh; calc_div.R)

1_BWA_index_ref.sh - Step 1 of variant calling pipeline: index reference genome

2_BWA_mem.sh - Step 2 of variant calling pipeline: map PE reads to reference genome

3_picard.sh - Step 3 of variant calling pipeline: quality check and mark mapped reads (sam/bam files)

4_GATK.sh - Step 4 of variant calling pipeline: realign bam file

5_GATK.sh - Step 5 of variant calling pipeline: HaplotypeCaller - call variants and record in variant call file (VCF)

6_GATK.sh - Step 6 of variant calling pipeline: combine files from both samples (desert and prairie) and extract only SNPs into new VCF

7_GATK.sh - Step 7 of variant calling pipeline: filter SNPs to selection criteria

8_snpEff.sh - Step 8 of variant calling pipeline: filter genic SNPs to those with highest likely impact (based on gene annotations)

9_neutral_filtering.sh - Step 9 of variant calling pipeline: filter intergenic SNPs to those farthest from any known genes (and therefore least likely to face linkage to sites under selection)

gatk_pipeline.sh - Steps 1 through 7 in variant calling pipeline above placed more cleanly in a single file as written for a single genome assembly. This script was not used for variant calling in our process, but is provided for better clarity/context.

format_conversion_scripts.R - R code converting genotypes to and from certain dataset formats for input into other programs and scripts

genepop_analysis.R - R code for analyses done in genepop

adegenet.R - R code for all analyses with Adegenet and Mantel Test (Figure 4; Figure B.2)

LEA_pop_structure.R - R code for LEA plots (Figure 5; Figure B.2)

zhangyimou.R - Seward Lee's code for plot color schemes following the films of Zhang Yimou. https://github.com/sewardlee337/zhangyimou/blob/master/R/palette_sample.R