Title

A reference genome assembly for the continentally distributed ring-necked snake, Diadophis punctatus

Journal

Journal of Heredity, 114(6)

ISSN

0022-1503

Authors

Westeen, Erin P
Escalona, Merly
Beraut, Eric
et al.

Publication Date

2023-11-15

DOI

10.1093/jhered/esad051

Peer reviewed

Journal of Heredity, 2023, **114**, 690–697
https://doi.org/10.1093/jhered/esad051
Advance access publication 9 September 2023
Genome Resources

American Genetic Association

OXFORD

Genome Resources

A reference genome assembly for the continentally distributed ring-necked snake, *Diadophis punctatus*

Erin P. Westeen[1,2], Merly Escalona[3], Eric Beraut[4], Mohan P.A. Marimuthu[5], Oanh Nguyen[5], Robert N. Fisher[6], Erin Toffelmier[7,8], H. Bradley Shaffer[7,8] and Ian J. Wang[1,2,*]

[1]Department of Environmental Science, Policy, and Management, University of California, Berkeley, Berkeley, CA 94720, United States,
[2]Museum of Vertebrate Zoology, University of California, Berkeley, Berkeley, CA 94720, United States,
[3]Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA 95064, United States,
[4]Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, Santa Cruz, CA 95064, United States,
[5]DNA Technologies and Expression Analysis Core Laboratory, Genome Center, University of California, Davis, Davis, CA 95616, United States,
[6]U.S. Geological Survey, Western Ecological Research Center, San Diego, CA, 92101, United States,
[7]Department of Ecology and Evolutionary Biology, University of California, Los Angeles, Los Angeles, CA 90095, United States,
[8]La Kretz Center for California Conservation Science, Institute of the Environment and Sustainability, University of California, Los Angeles, Los Angeles, CA 90095, United States

Address correspondence to I.J. Wang at the address above, ore-mail: ianwang@berkeley.edu

Corresponding Editor: Sara Ruane

Abstract

Snakes in the family Colubridae include more than 2,000 currently recognized species, and comprise roughly 75% of the global snake species diversity on Earth. For such a spectacular radiation, colubrid snakes remain poorly understood ecologically and genetically. Two subfamilies, Colubrinae (788 species) and Dipsadinae (833 species), comprise the bulk of colubrid species richness. Dipsadines are a speciose and diverse group of snakes that largely inhabit Central and South America, with a handful of small-body-size genera that have invaded North America. Among them, the ring-necked snake, *Diadophis punctatus*, has an incredibly broad distribution with 14 subspecies. Given its continental distribution and high degree of variation in coloration, diet, feeding ecology, and behavior, the ring-necked snake is an excellent species for the study of genetic diversity and trait evolution. Within California, six subspecies form a continuously distributed "ring species" around the Central Valley, while a seventh, the regal ring-necked snake, *Diadophis punctatus regalis* is a disjunct outlier and Species of Special Concern in the state. Here, we report a new reference genome assembly for the San Diego ring-necked snake, *D. p. similis*, as part of the California Conservation Genomics Project. This assembly comprises a total of 444 scaffolds spanning 1,783 Mb and has a contig N50 of 8.0 Mb, scaffold N50 of 83 Mb, and BUSCO completeness score of 94.5%. This reference genome will be a valuable resource for studies of the taxonomy, conservation, and evolution of the ring-necked snake across its broad, continental distribution.

Key words: California Conservation Genomics Project, CCGP, conservation genetics, Dipsadinae, reference genome, snake

Introduction

Though secretive in nature and rarely observed in the open, the ring-necked snake, *Diadophis punctatus*, has one of the largest ranges of any North American snake (Stebbins 2003). It occurs continuously along the east coast from southeastern Canada and the United States, west to the Great Lakes, south through the midwest to the Sierra Madre Occidental and Oriental of northern Mexico, and in isolated populations throughout the western United States (Stebbins and McGinnis 2012). Rangewide, 14 subspecies have been recognized based largely on the geographic distribution of differences in coloration and other morphological characters, with 7 of those occurring in California surrounding the Central Valley in the low-mid elevation oak/conifer woodland belt (Stebbins and McGinnis 2012; Fig. 1A). However, recent molecular evidence indicates that only three distinct lineages may be present in the California Floristic Province (CFP) and that they are geographically structured but do not align with current subspecies boundaries based on color pattern and morphology (Fontanella et al. 2021). Because subspecies remain the most common unit of conservation and management decisions for *D. punctatus*, we refer to the subspecies nomenclature hereafter, although more work is clearly needed to resolve the taxonomy of this system. The ring-necked snake is also one of only two genera in the largely neotropical family Dipsadinae to reach California (the other is the genus *Contia*; Pyron et al. 2013), making it of considerable biogeographic interest.

Ring-necked snakes are small—typically adult body length is 30 to 40 cm—and frequently encountered under rocks, logs, or other cover objects. Their dorsal coloration is a cryptic gray or olive green, while the ventral coloration is bright yellow–orange or red and varies geographically

Received April 12, 2023; Accepted September 7, 2023

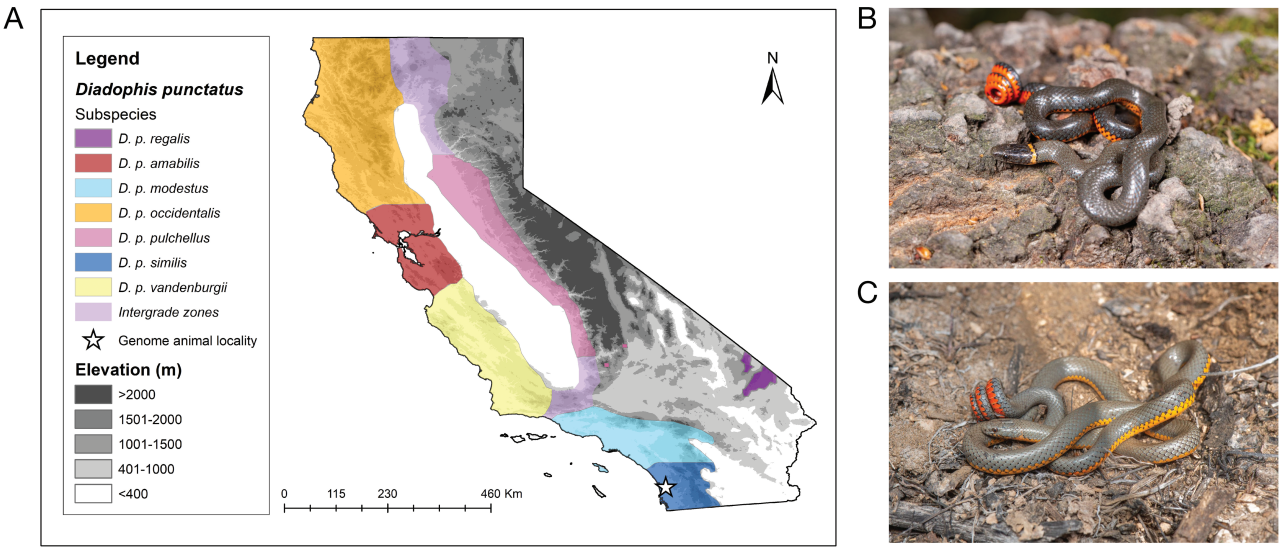

Fig. 1. (A) Range map of *Diadophis punctatus* in California showing the morphological subspecies boundaries as defined by Blanchard (1942), though recent work by Fontanella et al. (2021) found evidence for only three lineages in the state. (B) An individual from California showing the characteristic ring-neck coloration, bright red ventral coloration, and tail curling behavior (*D. p. amabilis* or Western California Lineage). (C) Another individual lacking characteristic the ring-neck coloration, displaying some of the variation in ventral coloration present in *D. punctatus*, and exhibiting tail curling behavior (*D. p. regalis*).

(Stebbins and McGinnis 2012; Fig. 1B and C). The scientific name (diadem = headband, ophis = snake) refers to the characteristic brightly colored ring around the neck (the neck ring is the same color as the belly), though some individuals lack this collar entirely (Holycross and Mitchell 2020; Fig. 1B and C). When threatened, ring-necked snakes often coil their tail to reveal the bright color underneath; this response varies geographically, and is more common in western subspecies (Ditmars 1908; Grinnell 1908; Woodbury 1928; Stebbins 2003); one study showed this response to be a result of tactile, but not visual, stimuli (Cox et al. 2021). It is generally regarded as an antipredator defensive behavior, although the exact mechanism of how it may function remains unresolved. Other antipredator behaviors include musk excretion, thrashing, and death feigning (Greene 1988). The ring-necked snake displays dietary variation across its range, with eastern subspecies focusing on invertebrates and small vertebrates (Fitch 1975) and western subspecies feeding largely on squamate reptiles and their eggs (Gehlbach 1974). Though typically ascribed to a clade of "non-venomous" snakes, ring-necks have enlarged rear fangs associated with Duvernoy's glands, and subspecies vary in both dental and venom phenotypes (Blanchard 1942; Taub 1967; Mackessy 2002; Westeen et al. 2020). Lab studies have shown the venom of *D. p. regalis* to be lethal to several neonate snakes on which it preys, including the red cornsnake, wandering garter snake, and western patch-nosed snake (Hill and Mackessy 2000), and the venom of *D. p. occidentalis* to be lethal to northwestern garter snakes (O'Donnell et al. 2007).

Given their small size and resultant high surface area-to-volume ratio, and reliance on small leaf litter prey, ring-necked snakes are particularly susceptible to climate warming. Three California subspecies are listed in the April 2023 edition of the California Natural Diversity Database (2023) Special Animals List. *Diadophis p. modestus* and *D. p. similis* are designated as Sensitive by the US Forest Service, while *D. p. regalis* is listed as a California Species of Special Concern (Thomson et al. 2016). These designations are based on population declines, limited distributions, and habitat loss. The remarkable variation in coloration, diet, venom, and associated rear-fanged phenotypes, behavior, and habitat across its range makes the ring-necked snake an excellent system for intraspecific comparative studies. Uncertainty regarding subspecies boundaries, ongoing population declines of multiple subspecies, and the unique phylogenetic and biogeographic positioning of this species highlight the value of additional genomic resources. Here, we present an assembled genome for *D. p. similis*, produced as part of the California Conservation Genomics Project (CCGP; Shaffer et al. 2022). Genetic evidence suggests that *D. p. similis* and *D. p. modestus*, both of which occur in southern California (United States) and northern Baja California (Mexico), form a single lineage and could be considered for species status if supported by additional data (Fontanella et al. 2021). In any case, we selected this subspecies for our *D. punctatus* genome assembly because the populations in southern California are sister to the northern California lineages that form the ring species structure (Fontanella et al. 2021), so it should provide a valuable reference for studying the evolution and biogeography of the ring species dynamic in this system. This annotated genome will serve as an invaluable resource for further study of this widespread, ecologically sensitive species.

Methods

Biological materials

We collected an adult female *D. p. similis* (collector field number HBS135679) in a small habitat patch just north of Camino Del Sur, about 400 m northeast of the intersection of Camino Del Sur and Bing Crosby Blvd, 4S Ranch, San Diego County California (33.0214, –117.1451) in December 2019 (CDFW entity permit no. SC-838 issued to the US Geological Survey). After several months in captivity, we removed and immediately flash froze liver tissue in liquid nitrogen and

stored it at −80 °C until extraction of genomic DNA (gDNA). After tissue harvesting, the specimen was formalin preserved and will be deposited in the Museum of Vertebrate Zoology at UC Berkeley.

HiFi library preparation and sequencing

High molecular weight (HMW) gDNA was extracted from 28 mg of liver tissue using the Nanobind Tissue Big DNA kit as per the manufacturer's instructions (Pacific BioSciences—PacBio, Menlo Park, CA). The DNA purity was estimated using absorbance ratios (260/280 = 1.83 and 260/230 = 2.10) on a NanoDrop ND-1000 spectrophotometer. The final DNA yield (195 ng/μL; 24 μg) was quantified using the Quantus Fluorometer (QuantiFluor ONE dsDNA Dye assay, Promega, Madison, WI). The size distribution of the HMW DNA was estimated using the Femto Pulse system (Agilent, Santa Clara, CA) which found that 66% of the fragments were 150 kb or more in length.

The HiFi SMRTbell library was constructed using the SMRTbell Express Template Prep Kit v2.0 (PacBio, Cat. #100-938-900) according to the manufacturer's instructions. HMW gDNA was sheared to a target DNA size distribution between 15 and 20 kb. The sheared gDNA was concentrated using 0.45× of AMPure PB beads (Pacific Biosciences—PacBio, Menlo Park, CA; Cat. #100-265-900) for the removal of single-strand overhangs at 37 °C for 15 min, followed by further enzymatic steps of DNA damage repair at 37 °C for 30 min, end repair and A-tailing at 20 °C for 10 min and 65 °C for 30 min, and ligation of overhang adapter v3 at 20 °C for 60 min and 65 °C for 10 min to inactivate the ligase. It was then nuclease treated at 37 °C for 1 h. The SMRTbell library was purified and concentrated with 0.45× Ampure PB beads (PacBio, Cat. #100-265-900) for size selection using the BluePippin/PippinHT system (Sage Science, Beverly, MA; Cat. #BLF7510/HPE7510) to collect fragments greater than 7 to 9 kb. The 15 to 20 kb average HiFi SMRTbell library was sequenced at UC Davis DNA Technologies Core (Davis, CA) using four 8M SMRT cells, Sequel II sequencing chemistry 2.0, and 30-h movies each on a PacBio Sequel II sequencer.

Omni-C library preparation and sequencing

The Omni-C library was prepared using the Dovetail Omni-C Kit (Dovetail Genomics, Scotts Valley, CA) according to the manufacturer's protocol with slight modifications. First, specimen tissue was thoroughly ground with a mortar and pestle while cooled with liquid nitrogen. Subsequently, chromatin was fixed in place in the nucleus. The suspended chromatin solution was then passed through 100 and 40 μm cell strainers to remove large debris. Fixed chromatin was digested under various conditions of DNase I until a suitable fragment length distribution of DNA molecules was obtained. Chromatin ends were repaired and ligated to a biotinylated bridge adapter followed by proximity ligation of adapter-containing ends. After proximity ligation, crosslinks were reversed and the DNA was purified from proteins. Purified DNA was treated to remove biotin that was not internal to ligated fragments. An NGS library was generated using an NEB Ultra II DNA Library Prep kit (New England Biolabs—NEB, Ipswich, MA) with an Illumina-compatible y-adaptor. Biotin-containing fragments were then captured using streptavidin beads. The postcapture product was split into two replicates prior to PCR enrichment to preserve library complexity with each replicate receiving unique dual indices. The library was sequenced at Vincent J. Coates Genomics Sequencing Lab (Berkeley, CA) on an Illumina NovaSeq platform (Illumina, San Diego, CA) to generate approximately 100 million 2 × 150 bp read pairs per GB of genome size.

Nuclear genome assembly

We assembled the *D. p. similis* genome following the CCGP assembly pipeline Version 4.0, which uses PacBio HiFi reads and Omni-C data to produce a high quality and highly contiguous assembly while minimizing manual curation (outlined in Table 1). We removed remnant adapter sequences from the PacBio HiFi dataset using HiFiAdapterFilt (Sim et al. 2022) and obtained the initial dual or partially phased diploid assembly (http://lh3.github.io/2021/10/10/introducing-dual-assembly) using HiFiasm (Cheng et al. 2021). We tagged output haplotype 1 as the primary assembly, and output haplotype 2 as the alternate assembly. We scaffolded both assemblies using the Omni-C data with SALSA (Ghurye et al. 2017, 2018). Next, we identified sequences corresponding to haplotypic duplications, contig overlaps, and repeats on the primary assembly with purge_dups [Version 1.2.5] (Guan et al. 2020) and transferred them to the alternate assembly.

We generated Omni-C contact maps for both assemblies by aligning the Omni-C data against the corresponding assembly with BWA-MEM (Li 2013), identified ligation junctions, and generated Omni-C pairs using pairtools (Goloborodko et al. 2018). We generated a multiresolution Omni-C matrix with cooler (Abdennur and Mirny 2020) and balanced it with hicExplorer (Ramírez et al. 2018). We used HiGlass (Kerpedjiev et al. 2018) and the PretextSuite (https://github.com/wtsi-hpag/PretextView; https://github.com/wtsi-hpag/PretextMap; https://github.com/wtsi-hpag/PretextSnapshot) to visualize the contact maps. We checked the contact maps for major misassemblies, cutting the scaffolds at the gaps where misassemblies were identified. No further joins were made after this step. Using the PacBio HiFi reads and YAGCloser (https://github.com/merlyescalona/yagcloser), we closed some of the remaining gaps generated during scaffolding. We then checked for contamination using the BlobToolKit Framework (Challis et al. 2020). Finally, we trimmed remnants of sequence adaptors and mitochondrial contamination identified during the contamination screening performed by NCBI.

Mitochondrial genome assembly

We assembled the mitochondrial genome of *D. p. similis* from the PacBio HiFi reads using the reference-guided pipeline MitoHiFi (https://github.com/marcelauliano/MitoHiFi; Allio et al. 2020). The mitochondrial sequence of *Hypsiglena jani jani* (NCBI:MT561500.1; Myers and Mulcahy 2020), another member of the family Dipsadinae, was used as the starting sequence. After completion of the nuclear genome, we searched for matches of the resulting mitochondrial assembly sequence in the nuclear genome assembly using BLAST+ (Camacho et al. 2009) and filtered out contigs and scaffolds from the nuclear genome with a percentage of sequence identity >99% and size smaller than the mitochondrial assembly sequence.

Table 1. Assembly pipeline and software usage. Software citations are listed in the text.

Assembly	Software and options §	Version
Filtering PacBio HiFi adapters	HiFiAdapterFilt	Commit 64d1c7b
K-mer counting	Meryl (k = 21)	1
Estimation of genome size and heterozygosity	GenomeScope	2
De novo *assembly (contiging)*	HiFiasm (Hi-C Mode, –primary, output p_ctg.hap1, p_ctg.hap2)	0.16.1-r375
Scaffolding		
Omni-C Scaffolding	SALSA (-DNASE, -i 20, -p yes)	2
Gap closing	YAGCloser (-mins 2 -f 20 -mcc 2 -prt 0.25 -eft 0.2 -pld 0.2)	Commit 0e34c3b
Omni-C Contact map generation		
Short-read alignment	BWA-MEM (-5SP)	0.7.17-r1188
SAM/BAM processing	Samtools	1.11
SAM/BAM filtering	Pairtools	0.3.0
Pairs indexing	Pairix	0.3.7
Matrix generation	Cooler	0.8.10
Matrix balancing	hicExplorer (hicCorrectmatrix correct --filterThreshold -2 4)	3.6
Contact map visualization	HiGlass	2.1.11
	PretextMap	0.1.4
	PretextView	0.1.5
	PretextSnapshot	0.0.3
Organelle assembly		
Mitogenome assembly	MitoHiFi (-r, -p 50, -o 1)	Commit c06ed3e
Genome quality assessment		
Basic assembly metrics	QUAST (--est-ref-size)	5.0.2
Assembly completeness	BUSCO (-m geno, -l tetrapoda)	5.0.0
	Merqury	2020-01-29
Contamination screening		
Local alignment tool	BLAST+	2.1
General contamination screening	BlobToolKit	2.3.3

Genome size estimation and quality assessment

We generated k-mer counts from the PacBio HiFi reads using meryl (https://github.com/marbl/meryl). The k-mer database was then used in GenomeScope2.0 (Ranallo-Benavidez et al. 2020) to estimate genome features including genome size, heterozygosity, and repeat content. To obtain general contiguity metrics, we ran QUAST (Gurevich et al. 2013). To evaluate genome quality and completeness we used BUSCO (Manni et al. 2021) with the tetrapoda ortholog database (tetrapoda_odb10) which contains 5,310 genes. Assessment of base level accuracy (QV) and k-mer completeness was performed using the previously generated meryl database and merqury (Rhie et al. 2020). We further estimated genome assembly accuracy via BUSCO gene set frameshift analysis using the pipeline described in Korlach et al. (2017).

Measurements of the size of the phased blocks is based on the size of the contigs generated by HiFiasm on HiC mode. We follow the quality metric nomenclature established by Rhie et al. (2021), with the genome quality code $x{\cdot}y{\cdot}P{\cdot}Q{\cdot}C$, where, $x = \log_{10}[\text{contig NG50}]$; $y = \log_{10}[\text{scaffold NG50}]$; $P = \log_{10}[\text{phased block NG50}]$; Q = Phred base accuracy QV (quality value); C = % genome represented by the first "n" scaffolds, following a known karyotype of 2n = 36 for *D. p. occidentalis* (Bury et al. 1970). Quality metrics for the notation were calculated on the primary assembly.

Results

The Omni-C and PacBio HiFi sequencing libraries generated 283.9 million read pairs and 4.2 million reads, respectively. The latter yielded 48.16-fold coverage (N50 read length 18,494 bp; minimum read length 44 bp; mean read length 18,061 bp; maximum read length 63,713 bp) based on the Genomescope 2.0 genome size estimation of 1.5 Gb. Based on PacBio HiFi reads, we estimated 0.166% sequencing error rate and 0.977% nucleotide heterozygosity rate. The k-mer spectrum based on PacBio HiFi reads show a bimodal distribution with two major peaks at ~23- and ~45-fold coverage (Fig. 2A), where peaks correspond to homozygous and heterozygous states of a diploid species. The distribution presented in this k-mer spectrum supports that of a low heterozygosity profile.

The final assembly (rDiaPun1) consists of two pseudo haplotypes, primary and alternate. The genome size of the haplotypes is similar but not identical to the value estimated by Genomescope2.0 (Fig. 2A). The primary assembly consists of 444 scaffolds spanning 1.78 Gb with contig N50 of 8 Mb, scaffold N50 of 83.6 Mb, largest contig of 57.6 Mb, and largest scaffold of 321.1 Mb. The alternate assembly consists of 2,064 scaffolds, spanning 1.57 Gb with contig N50 of 6.58 Mb, scaffold N50 of 65.4 Mb, largest contig of 46 Mb, and largest scaffold of 185.2 Mb. Assembly statistics are reported in tabular form in Table 2, and graphically for the primary assembly in Fig. 2B.

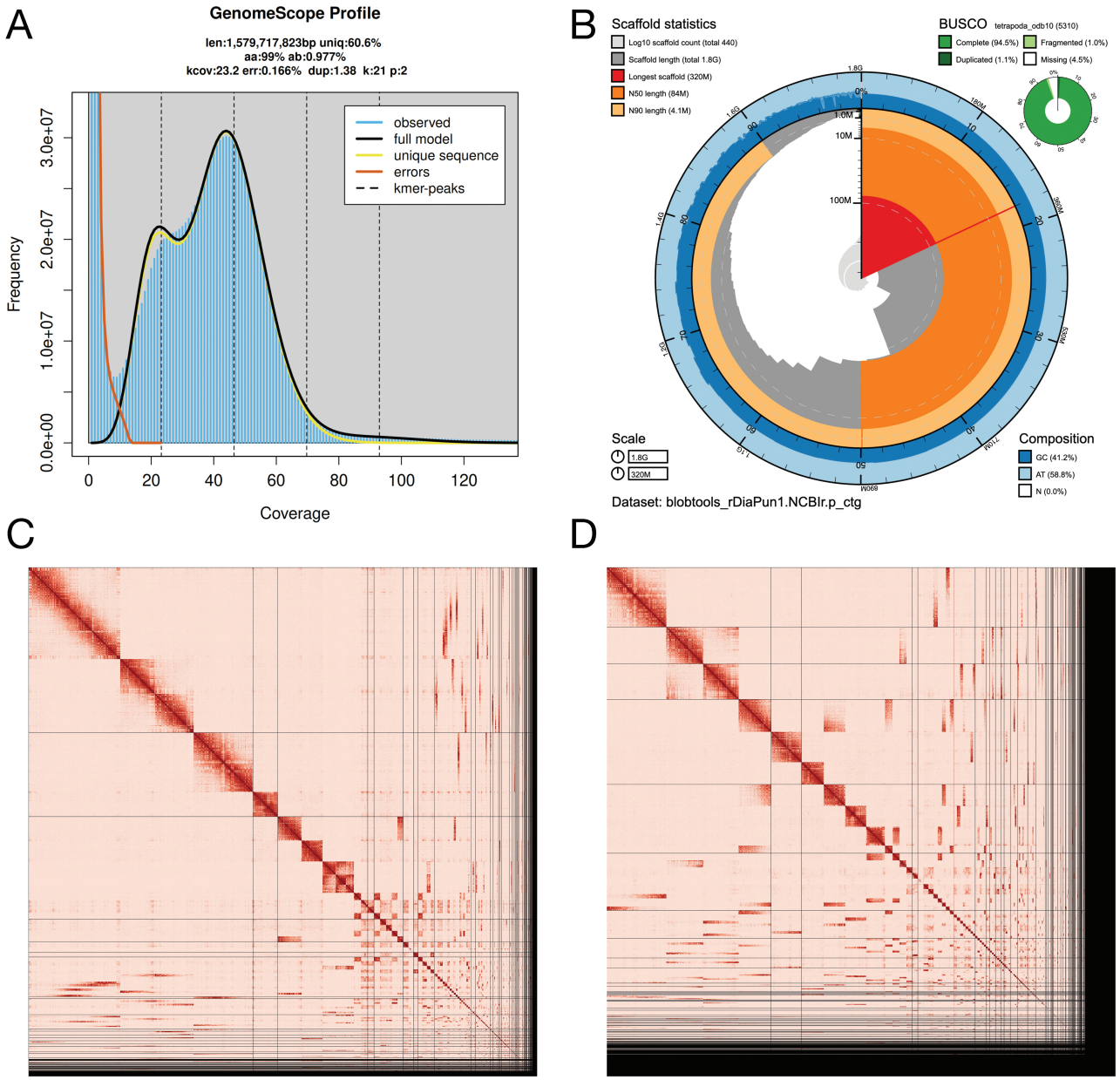

Fig. 2. Visual overview of genome assembly metrics. (A) K-mer spectra output generated from PacBio HiFi data without adapters using GenomeScope2.0. The bimodal pattern observed corresponds to a diploid genome. K-mers covered at lower coverage and lower frequency correspond to differences between haplotypes, whereas the higher coverage and higher frequency k-mers correspond to the similarities between haplotypes. (B) BlobToolKit Snail plot showing a graphical representation of the quality metrics presented in Table 2 for the *Diadophis punctatus* primary assembly (rDiaPun1). The plot circle represents the full size of the assembly. From the inside out, the central plot covers length-related metrics. The red line represents the size of the longest scaffold; all other scaffolds are arranged in size order moving clockwise around the plot and drawn in gray starting from the outside of the central plot. Dark and light orange arcs show the scaffold N50 and scaffold N90 values. The central light gray spiral shows the cumulative scaffold count with a white line at each order of magnitude. White regions in this area reflect the proportion of Ns in the assembly. The dark vs. light blue area around it shows mean, maximum, and minimum GC vs. AT content at 0.1% intervals (Challis et al. 2020). (C and D) Omni-C Contact maps for the primary (2C) and alternate (2D) genome assembly generated with PretextSnapshot. Omni-C contact maps translate proximity of genomic regions in 3-D space to contiguous linear organization. Each cell in the contact map corresponds to sequencing data supporting the linkage (or join) between two of such regions. Scaffolds are separated by black lines and higher density corresponds to higher levels of fragmentation.

We identified a total of 17 misassemblies, 10 on the primary assembly and 7 on the alternate, and broke the corresponding joins made by SALSA on both assemblies. We were able to close a total of 32 gaps, 25 on the primary and 7 on the alternate assembly. We further filtered out 3 contigs corresponding to arthropod contaminants (1 contig from the primary assembly and 2 from the alternate). Finally, we filtered out 3 contigs (1 from the primary and 2 from the alternate) corresponding to mitochondrial contamination. No further contigs were removed. The primary assembly has a BUSCO completeness score of 94.5% using the Tetrapoda gene set, a per base quality (QV) of 60.71, a k-mer completeness of 93.15, and a frameshift indel QV of 47.51. The alternate assembly has a BUSCO completeness score of 79.6% again using the Tetrapoda gene set, a per base quality (QV) of 59.57, a k-mer completeness of 77.93, and a frameshift indel

QV of 47.13. We have deposited scaffolds corresponding to both primary and alternate haplotype (see Table 2 and Data Availability for details).

We assembled a mitochondrial genome with MitoHiFi. Final mitochondrial genome size was 17,158 bp. The base composition of the final assembly version is $A = 33.14\%$, $C = 28.05\%$, $G = 13.33\%$, $T = 25.48\%$, and consists of 22 unique transfer RNAs and 13 protein-coding genes.

Discussion

Of the 45 snake genomes currently available via GenBank, only two other members of Dipsadinae are represented. The blunt-headed tree snake, *Imantodes cenchoa*, is the species with an annotated genome currently most closely related to the ring-necked snake, though the two species are separated by an estimated 30 My of evolutionary history (Pyron et al. 2013). The two genomes are very similar in size, with

Table 2. Sequencing and assembly statistics and accession numbers. The assembly quality code is provided in the form *x·y·P·Q·C*, where, $x = \log_{10}$[contig NG50]; $y = \log_{10}$[scaffold NG50]; $P = \log_{10}$[phased block NG50]; Q = Phred base accuracy QV (quality value); C = % genome represented by the first "n" scaffolds, following a known karyotype 2n = 36 for *D. punctatus* (Bury et al. 1970). BUSCO scores are presented for the percentages of (C)omplete, (S)ingle, (D)uplicated, (F)ragmented, and (M)issing genes for the number of BUSCO genes (n = 5,310) in the set/database. Read coverage and NGx statistics (§) have been calculated based on a genome size of 1.5 Gb. Genome assembly quality metrics, accession numbers and BUSCO scores are presented for the P(rimary) and (A)lternate assembly values.

Bio Projects & Vouchers	CCGP NCBI BioProject					PRJNA720569	
	Genera NCBI BioProject					PRJNA765811	
	Species NCBI BioProject					PRJNA808336	
	NCBI BioSample					SAMN25872410	
	Specimen identification					HBS135679	
	NCBI Genome accessions					Primary	Alternate
	Assembly accession					JALIGW000000000	JALIGX000000000
	Genome sequences					GCA_023053685.1	GCA_023053665.1
Genome Sequence	PacBio HiFi reads	Run				1 PACBIO_SMRT (Sequel II) run: 4.2M spots, 76.1G bases, 56.3Gb	
		Accession				SRX15303501	
	Omni-C Illumina reads	Run				2 ILLUMINA (Illumina NovaSeq 6000) runs: 284M spots, 85.8G bases, 28.4Gb	
		Accession				SRX15303502, SRX15303503	
Genome Assembly Quality Metrics	Assembly identifier (Quality code)					rDiaPun1(6.7.P7.Q60.C74)	
	HiFi Read coverage §					38.04X	
						Primary	Alternate
	Number of contigs					1,167	2,636
	Contig N50 (bp)					8,013,852	6,581,262
	Contig NG50 §					10,096,176	6,581,262
	Longest Contigs					57,679,849	46,002,525
	Number of scaffolds					444	2,064
	Scaffold N50 (bp)					83,654,930	65,413,798
	Scaffold NG50 §					86,669,289	65,413,798
	Largest scaffold					321,190,879	185,256,169
	Size of final assembly (bp)					1,783,023,707	1,577,714,162
	Phased block NG50 §					9,980,834	8,012,668
	Gaps per Gbp					405 (723)	362 (572)
	Indel QV (Frame shift)					47.51674281	47.13028282
	Base pair QV					60.7177	59.5778
						Full assembly = 60.1452	
	k-mer completeness					93.1582	77.9347
						Full assembly = 98.5205	
	BUSCO completeness		C	S	D	F	M
	(tetrapoda) n = 5,310	P	94.50%	93.40%	1.10%	1.00%	4.50%
		A	79.60%	78.90%	0.70%	1.20%	19.20%
	Organelles		1 Partial mitochondrial sequence			JALIGW010000444.1	

the *Imantodes* genome (1.4 Gb) slightly smaller than our ring-necked snake assembly at 1.5 Gb. The other snakes represented span 10 families and range in size from 1.127 to 2.038 Gb. The largest squamate genome assembly size is currently 2.856 Gb (for the western fence lizard, *Sceloporus occidentalis*; Bishop et al. 2023). Among the currently available snake genome assemblies, comprising 10 families, contig N50 values range from 0.8 kb to 54.1 Mb to scaffold N50 values from 1.5 kb to 266 Mb. This *D. p. similis* genome assembly, with a contig N50 of 8 Mb and scaffold N50 of 83.7 Mb (Table 2), is well within the range of scaffold lengths of these other existing genomes.

The *D. p. similis* reference assembly is one of seven species of wide-ranging California snakes that are being produced by the CCGP, including the glossy snake, *Arizona elegans* (Wood et al. 2022), rubber boa, *Charina bottae* (Grismer et al. 2022), and southern Pacific rattlesnake (Westeen et al. 2023). This genome provides an important phylogenetic branch to the diversity of California with CCGP resources now available (Toffelmier et al. 2022), as well as being one of the few members of the diverse subfamily Dipsadinae with a reference-level genome. It also will provide a new resource for studies regarding the evolutionary history, diversification, and conservation of *D. punctatus*. Recent findings suggest that only three genetically distinct clades of *Diadophis* are present in the CFP (Fontanella et al. 2021), which contrasts with the original subspecies descriptions by Blanchard (1942). A complete understanding of lineage relationships and boundaries within *D. punctatus* will inform conservation and management strategies going forward.

Acknowledgments

PacBio Sequel II library prep and sequencing were carried out at the DNA Technologies and Expression Analysis Cores at the UC Davis Genome Center, supported by NIH Shared Instrumentation Grant 1S10OD010786-01. Deep sequencing of Omni-C libraries used the Novaseq S4 sequencing platforms at the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 OD018174 Instrumentation Grant. We thank the staff at the UC Davis DNA Technologies and Expression Analysis Cores and the UC Santa Cruz Paleogenomics Laboratory for their diligence and dedication to generating high-quality sequence data. Partial support was provided by Illumina for Omni-C sequencing. We also thank the California Department of Fish and Wildlife (SC-838) for granting scientific research permit for tissue collection. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the US Government.

Funding

This work was supported by the California Conservation Genomics Project, with funding provided to the University of California by the State of California, State Budget Act of 2019 [UC Award ID RSI-19-690224].

Data Availability

Data generated for this study are available under NCBI BioProject PRJNA808336. Raw sequencing data for sample HBS135679 (NCBI BioSample SAMN25872410) are deposited in the NCBI Short Read Archive (SRA) under SRX15303501 for PacBio HiFi sequencing data, and SRX15303502, SRX15303503 for the Omni-C Illumina sequencing data. GenBank accessions for both primary and alternate assemblies are GCA_023053685.1 and GCA_023053665.1; and for genome sequences JJALIGW000000000 and JALIGX000000000. The GenBank organelle genome assembly for the mitochondrial genome is JALIGW010000444.1. Assembly scripts and other data for the analyses presented can be found at the following GitHub repository: www.github.com/ccgproject/ccgp_assembly.

References

Abdennur N, Mirny LA. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics*. 2020:36:311–316. https://doi.org/10.1093/bioinformatics/btz540.

Allio R, Schomaker-Bastos A, Romiguier J, Prosdocimi F, Nabholz B, Delsuc F. MitoFinder: efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. *Mol Ecol Resour*. 2020:20:892–905. https://doi.org/10.1111/1755-0998.13160.

Bishop AP, Westeen EP, Yuan ML, Escalona M, Beraut E, Fairbairn C, Marimuthu MPA, Nguyen O, Chumchim N, Toffelmier E, et al. Assembly of the largest squamate reference genome to date: the western fence lizard, Sceloporus occidentalis. *J Hered*. 2023:114:521–528. https://doi.org/10.1093/jhered/esad037.

Blanchard F. The ring-neck snakes, genus Diadophis. *Bull Chicago Acad Sci*. 1942:7:1–144.

Bury B, Gress F, Gorman GC. Karyotypic survey of some colubrid snakes from western North America. *Herpetologica*. 1970:26:461–466.

California Natural Diversity Database (CNDDB). *Special animals list*. Sacramento (CA): California Department of Fish and Wildlife; 2023.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. *BMC Bioinf*. 2009:10:421.

Challis R, Richards E, Rajan J, Cochrane G, Blaxter M. BlobToolKit – interactive quality assessment of genome assemblies. *G3 Genes Genomes Genet*. 2020:10:1361–1374. https://doi.org/10.1534/g3.119.400908.

Cheng H, Jarvis ED, Fedrigo O, Koepfli K-P, Urban L, Gemmell NJ, Heng L. *Robust haplotype-resolved assembly of diploid individuals without parental data*. https://doi.org/10.48550/arXiv.2109.04785, 2021, preprint: not peer reviewed.

Cox CL, Chung AK, Blackwell C, Davis MM, Gulsby M, Islam H, Miller N, Lambert C, Lewis O, Rector IV, et al. Tactile stimuli induce deimatic antipredator displays in ringneck snakes. *Ethology*. 2021:127:465–474.

Ditmars R. *The reptile book; a comprehensive, popularised work on the structure and habits of the turtles, tortoises, crocodilians, lizards and snakes which inhabit the United States and northern Mexico*. New York (NY): Doubleday, Page & Company; 1908.

Fitch HS. A demographic study of the ringneck snake (Diadophis punctatus) in Kansas. *Univ Kansas Mus Nat Hist Misc Pub*. 1975:62:1–53.

Fontanella FM, Miles E, Strott P. Integrated analysis of the ringneck snake *Diadophis punctatus* complex (Colubridae: Dipsadidae) in a biodiversity hotspot provides the foundation for conservation reassessment. *Biol J Linn Soc*. 2021:133:105–119.

Gehlbach FR. Evolutionary relations of southwestern ringneck snakes (*Diadophis punctatus*). *Herpetologica*. 1974:30:140–148.

Ghurye J, Pop M, Koren S, Bickhart D, Chin C-S. Scaffolding of long read assemblies using long range contact information. *BMC Genomics*. 2017:18:527. https://doi.org/10.1186/s12864-017-3879-z.

Ghurye J, Rhie A, Walenz, BP, Schmitt A, Selvaraj S, Pop M, Phillippy AM, Koren S. Integrating Hi-C links with assembly graphs for

chromosome-scale assembly. *PLoS Computational Biology*. 2018:15:e1007273. https://doi.org/10.1371/journal.pcbi.1007273.
Goloborodko A, Abdennur N, Venev S, Brandão, Gfudenberg. mirnylab/pairtools: v0.2.0. 2018. https://doi.org/10.5281/zenodo.1490831.
Greene HW. Antipredator mechanisms in reptiles. In: Gans C, Huey RB, editors. *Biology of the reptilians*. New York (NY): Alan R. Liss; 1988. p. 1–152.
Grinnell J. The biota of the San Bernardino Mountains. Berkeley, CA, USA: University of California Press; 1908:5:1–170.
Grismer JL, Escalona MC, Miller C, Beraut E, Fairbairn CW, Marimuthu MPA, Nguyen O, Toffelmier E, Wang IJ, Shaffer HB. Reference genome of the rubber boa, *Charina bottae* (Serpentes: Boidae). *J Hered*. 2022:113:641–648.
Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*. 2020:36:2896–2898. https://doi.org/10.1093/bioinformatics/btaa025.
Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. 2013:29:1072–1075. https://doi.org/10.1093/bioinformatics/btt086.
Hill RE, Mackessy SP. Characterization of venom (Duvernoy's secretion) from twelve species of colubrid snakes and partial sequence of four venom proteins. *Toxicon*. 2000:38:1663–1687.
Holycross AT, Mitchell JC. *Snakes of Arizona*. Rodeo (NM): ECO Publishing; 2020.
Kerpedjiev P, Abdennur N, Lekschas F, McCallum C, Dinkla K, Strobelt H, Luber JM, Oulette SB, Azhir A, Kumar N, et al. HiGlass: web-based visual exploration and analysis of genome interaction maps. *Genome Biol*. 2018:19:125. https://doi.org/10.1186/s13059-018-1486-1.
Korlach J, Gedman G, Kingan SB, Chin C-S, Howard JT, Audet J-N, Cantin L, Jarvis ED. De novo PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. *GigaScience*. 2017:6:1–16. https://doi.org/10.1093/gigascience/gix085.
Li H. *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. https://doi.org/10.48550/arXiv:1303.3997, 2013, preprint: not peer reviewed.
Mackessy SP. Biochemistry and pharmacology of colubrid snake venoms. *J Toxicol Toxin Rev*. 2002:21:43–83.
Manni M, Berkeley MR, Seppey M, Simao FA, Zdobnov EM. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol*. 2021:38:4647–4654.
Myers EA, Mulcahy DG. Six additional mitochondrial genomes for North American nightsnakes (Dipsadidae: *Hypsiglena*) and a novel gene feature for advanced snakes. *Mitochondrial DNA B Resour*. 2020:5:3056–3058. https://doi.org/10.1080/23802359.2020.1797573.
O'Donnell RP, Staniland K, Mason RT. Experimental evidence that oral secretions of northwestern ring-necked snakes (*Diadophis punctatus occidentalis*) are toxic to their prey. *Toxicon*. 2007:50:810–815.
Pyron RA, Burbrink FT, Wiens JJ. A phylogeny and revised classification of Squamata, including 4161 species of lizards and snakes. *BMC Evol Biol*. 2013:13:93.
Ramírez F, Bhardwaj V, Arrigoni L, Lam KC, Grüning BA, Villaveces J, Habermann B, Akhtar A, Manke T. High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat Commun*. 2018:9:189. https://doi.org/10.1038/s41467-017-02525-w.
Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun*. 2020:11:1432. https://doi.org/10.1038/s41467-020-14998-3.
Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W, Fungtammasan A, Kim J, et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature*. 2021:592:737–746. https://doi.org/10.1038/s41586-021-03451-0.
Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol*. 2020:21:245. https://doi.org/10.1186/s13059-020-02134-9.
Shaffer HB, Toffelmier E, Corbett-Detig RB, Escalona M, Erickson B, Fiedler P, Gold M, Harrigan RJ, Hodges S, Luckau TK, et al. Landscape genomics to enable conservation actions: the California Conservation Genomics Project. *J Hered*. 2022:113:577–588. https://doi.org/10.1093/jhered/esac020.
Sim SB, Corpuz RL, Simmonds TJ, Geib SM. HiFiAdapterFilt, a memory efficient read processing pipeline, prevents occurrence of adapter sequence in PacBio HiFi reads and their negative impacts on genome assembly. *BMC Genomics*. 2022:23:157. https://doi.org/10.1186/s12864-022-08375-1.
Stebbins R. *Western reptiles and amphibians*. New York (NY): Houghton Mifflin; 2003.
Stebbins R, McGinnis S. *Field guide to amphibians and reptiles of California*. Berkeley and Los Angeles (CA): University of California Press; 2012.
Taub AM. Comparative histological studies on Duvernoy's gland of colubrid snakes. *Bull Am Mus Nat Hist*. 1967:138:1–50.
Thomson RC, Wright AN, Shaffer HB. *California amphibian and reptile species of special concern*. Oakland, CA: University of California Press. 390 + xv pages; 2016.
Toffelmier E, Beninde J, Shaffer HB. The phylogeny of California, and how it informs setting multi-species conservation priorities. *J Hered*. 2022:113:597–603.
Westeen EP, Durso AM, Grundler MC, Rabosky DL, Davis Rabosky AR. What makes a fang? Phylogenetic and ecological controls on tooth evolution in rear-fanged snakes. *BMC Evol Biol*. 2020:20:80. https://doi.org/10.1186/s12862-020-01645-0.
Westeen EP, Escalona EP, Holding ML, Beraut E, Fairbairn C, Marimuthu MPA, Nguyen O, Perri R, Fisher RN, Toffelmier E, et al. A genome assembly for the southern Pacific rattlesnake, *Crotalus oreganus helleri*, in the western rattlesnake species complex. *J Hered*. 2023:2023:esad045. https://doi.org/10.1093/jhered/esad045.
Wood DA, Richmond JQ, Escalona M, Marimuthu MPA, Nguyen O, Sacco S, Beraut E, Westphal M, Fisher RN, Vandergast AG, et al. Reference genome of the California glossy snake, *Arizona elegans occidentalis*: a declining California Species of Special Concern. *J Hered*. 2022:113:632–640.
Woodbury AM. The reptiles of Zion National Park. *Copeia*. 1928:16:14–21.