*Journal of Heredity*, 2022, **113**, 641–648
https://doi.org/10.1093/jhered/esac048
Advance access publication 3 September 2022
**Genome Resources**

American Genetic Association

OXFORD

Genome Resources

# Reference genome of the rubber boa, *Charina bottae* (Serpentes: Boidae)

**Jesse L. Grismer[1], Merly Escalona[2], Courtney Miller[3], Eric Beraut[4], Colin W. Fairbairn[4], Mohan P.A. Marimuthu[5], Oanh Nguyen[5], Erin Toffelmier[3], Ian J. Wang[6,7], H. Bradley Shaffer[3,8]**

[1]Department of Biology, La Sierra University, Riverside, CA, United States,
[2]Department of Biomolecular Engineering, University of California, Santa Cruz, Santa Cruz, CA, United States,
[3]Department of Ecology & Evolutionary Biology, University of California, Los Angeles, Los Angeles, CA, United States,
[4]Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, Santa Cruz, CA, United States,
[5]DNA Technologies and Expression Analysis Core Laboratory, Genome Center, University of California, Davis, Davis, CA, United States,
[6]Department of Environmental Science, Policy, and Management, University of California, Berkeley, Berkeley, CA, United States,
[7]Museum of Vertebrate Zoology, University of California, Berkeley, Berkeley, CA, United States,
[8]La Kretz Center for California Conservation Science, Institute of the Environment and Sustainability, University of California, Los Angeles, Los Angeles, CA, United States

Address correspondence to J.L. Grismer at the address above, or e-mail: jgrismer@lasierra.edu.

Corresponding Editor: Rachel Meyer

## Abstract

The rubber boa, *Charina bottae* is a semi-fossorial, cold-temperature adapted snake that ranges across the wetter and cooler ecoregions of the California Floristic Province. The rubber boa is 1 of 2 species in the family Boidae native to California and currently has 2 recognized subspecies, the Northern rubber boa *C. bottae bottae* and the Southern rubber boa *C. bottae umbratica*. Recent genomic work on *C. bottae* indicates that these 2 subspecies are collectively composed of 4 divergent lineages that separated during the late Miocene. Analysis of habitat suitability indicates that *C. bottae umbratica* montane sky-island populations from southern California will lose the majority of their habit over the next 70 yr, and is listed as Threatened under the California Endangered Species Act. Here, we report a new, chromosome-level assembly of *C. bottae bottae* as part of the California Conservation Genomics Project (CCGP). Consistent with the reference genome strategy of the CCGP, we used Pacific Biosciences HiFi long reads and Hi-C chromatin-proximity sequencing technology to produce a de novo assembled genome. The assembly comprises 289 scaffolds covering 1,804,944,895 bp, has a contig N50 of 37.3 Mb, a scaffold N50 of 97 Mb, and BUSCO completeness score of 96.3%, and represents the first reference genome for the Boidae snake family. This genome will enable studies of genetic differentiation and connectivity among *C. bottae bottae* and *C. bottae umbratica* populations across California and help manage locally endemic lineages as they confront challenges from human-induced climate warming, droughts, and wildfires across California.

**Key words:** Boidae, California Conservation Genomics Project, CCGP, Charininae, conservation genetics, sky-island

## Introduction

The rubber boa, *Charina bottae* (Blainville 1835), is a cold-temperature and moisture-adapted semi-fossorial snake species that ranges from British Columbia, Canada to southern California in the southwest, and eastward to Utah, Wyoming, and Montana (Stewart 1977; Stebbins 2003). Currently, there are 2 subspecies that are generally recognized within *Charina* (Fig. 1), the Northern rubber boa *C. bottae bottae* (Van Denburgh 1920) and the Southern rubber boa *C. bottae umbratica* (Klauber 1943), the latter of which is listed as threatened under the California Endangered Species Act (Stewart 1977; Rodríguez-Robles et al. 2001; CDFG 2005; Grismer et al. 2022). A recent range-wide RADseq analysis recovered 4 deeply divergent geographic lineages within *Charina* that are confined to the Pacific Northwest and Great Basin (PGB), coastal California, the Sierra Nevada Mountains, and southern California (Grismer et al. 2022). *C. bottae bottae* is currently composed of the Coastal California, Sierra Nevada Mountains, and PGB lineages, while *C. bottae umbratica* contains multiple fragmented montane sky-island populations representing the Southern California lineage (Rodríguez-Robles et al. 2001; Grismer et al. 2022). These *C. bottae umbratica* mountain isolates occur in alpine habitats exclusively above 1,800 m in elevation and range from Breckenridge Mountain in eastern Kern County, CA southeastward to the San Jacinto Mountains in Riverside County (Rodríguez-Robles et al. 2001; Grismer et al. 2022).

There are many co-distributed cold-temperature and moisture-adapted species with varying levels of genetic substructuring that occur across the same wetter microhabitats

Received August 2, 2022; Accepted September 2, 2022


**Fig. 1.** (A) Northern rubber boa *Charina bottae bottae*. (B) Southern rubber boa, *Charina bottae umbratica*. (C) Rocky alpine habitats of *Charina*. (D) Distribution of the 2 *Charina* subspecies.

and ecoregions of California. Unfortunately, their natural histories and ecological requirements render them particularly vulnerable to the effects of future climate change (Minnich and Franco-Vizcaíno 2005; Vandgergast et al. 2008; Devit et al. 2013; Grismer et al. 2022). Since 1972 there has been an increase in annual temperatures, and droughts, and wildfire frequency across California stemming from human-induced warming (Abatzoglou et al. 2016; Westerling 2018; Williams et al. 2019). Many of these threats have heavily impacted forested and montane ecosystems, and species models of future habitat stability indicates that many *C. bottae umbratica* populations and other sky-island species southern California that could lose the majority of suitable habitat over the next 70 yr (Devitt et al. 2013; Williams et al. 2019; Grismer et al. 2022). In southern California, these montane sky-islands are the only areas that support these cold-temperature and moisture-adapted species, and represent a unique ecosystem in the region. Unfortunately, climate-related threats have the potential to impact their future integrity and the habitat specialists that in habiting them (Brehme et al. 2011; Syphard et al. 2016; Tracey et al. 2018).

Here, we report the first chromosome-level genome assembly for *C. bottae*, sequenced and assembled as part of the California Conservation Genomics Project (CCGP) (Shaffer et al. 2022). This is the third of 15 reptiles, and second of 7 snake species reference genomes constructed for the CCGP (Todd et al. 2022; Wood et al. 2022). This genome assembly will provide the first genome for the family Boidae and the Charininae subfamily. This genome will be a foundational resource for future studies on the conservation, ecophysiology, biogeography, and taxonomy of *C. bottae*.

## Methods

### Biological materials

An adult male *C. bottae bottae* was collected on 11/25/2021 (HBS 135849: HB Shaffer permit SCP 2480) from Skyline Boulevard in San Mateo County, CA (37.492778, –122.366528). Liver, skeletal muscle, heart, brain, intestine, testis, and kidney tissues were harvested, and flash frozen, and the specimen was formalin fixed and will be deposited at the

Museum of Vertebrate Zoology at the University of California, Berkeley. This sample is from the Coastal California lineage of *C. bottae bottae* which contains the type locality for *C. bottae*.

## Nucleic acid library extraction

We extracted high molecular weight (HMW) DNA from 38 mg of liver tissue (HBS135849) using the Nanobind Tissue Big DNA kit (Pacific BioSciences), following the manufacturer's instructions. We assessed DNA purity using absorbance ratios (260/280 = 1.84 and 260/230 = 2.34) on a NanoDrop ND-1000 spectrophotometer. We quantified DNA yield (234 ng/µl; 45.6 µg total) using a Quantus Fluorometer (QuantiFluor ONE dsDNA Dye assay, Promega). We estimated the size distribution of the HMW DNA using the Femto Pulse system (Agilent) and found that >86% of the DNA fragments were >125 kb.

## Nucleic acid library preparation

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, ligation of overhang adapter v3 at 20 °C for 60 min and 65 °C for 10 min to inactivate the ligase, 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 2 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, 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 purified from proteins. Purified DNA was treated to remove biotin that was not internal to ligated fragments, and an NGS library was generated using an NEB Ultra II DNA Library Prep kit (NEB, Ipswich, MA) with an Illumina compatible y-adaptor. Biotin-containing fragments were then captured using streptavidin beads, and the postcapture product was split into 2 replicates prior to PCR enrichment to preserve library complexity with each replicate receiving unique dual indices. The library was sequenced on an Illumina NovaSeq platform to generate approximately 100 million 2 × 150 bp read pairs per GB genome size.

## Nuclear genome assembly

We assembled the rubber boa genome following the CCGP assembly pipeline Version 4.0, as outlined in Table 1. As with other CCGP assemblies, our goal is to produce a high-quality and highly contiguous assembly using PacBio HiFi reads and Omni-C data while minimizing manual curation. We removed remnant adapter sequences from the PacBio HiFi dataset using HiFiAdapterFilt (Sim et al. 2022) and obtained the initial a 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, 2019).

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) and visualized the contact maps with 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, and manually cut the assemblies at the gaps where misassemblies were found. 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 *C. bottae* from the PacBio HiFi reads using the reference-guided pipeline MitoHiFi [Version 2] (https://github.com/marcelauliano/MitoHiFi; Allio et al. 2020). The mitochondrial sequence of *Eryx tataricus* (NCBI:MN646174.1) was used as the starting reference 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.

## 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 (Simão

**Table 1.** Assembly pipeline and software used.

| Assembly | Software and options[a] | 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) | 2 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 |

Software citations are listed in the text.
[a]Options detailed for nondefault parameters.

et al. 2015; Seppey et al. 2019; 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 are based on the size of the contigs generated by HiFiasm on HiC mode (initial diploid assembly). 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 $x$ = log10[contig NG50]; $y$ = log10[scaffold NG50]; $P$ = log10[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 *C. bottae* (Gorman and Gress 1970). Quality metrics for the notation were calculated on the primary assembly.

## Results

The Omni-C and PacBio HiFi sequencing libraries generated 123.5 million read pairs and 3.5 million reads, respectively. The latter yielded ~30-fold coverage (N50 read length 15,171 bp; minimum read length 71 bp; mean read length 15,089 bp; maximum read length of 48,995 bp) based on the GenomeScope 2.0 genome size estimation of 1.7 Gb. Based on PacBio HiFi reads, we estimated 0.141% sequencing error rate and 0.221% nucleotide heterozygosity rate. The k-mer spectrum based on PacBio HiFi reads (Fig. 2A) show a bimodal distribution with 2 major peaks at 15- and 30-fold coverage, 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 (rChaBot1) consists of 2 pseudo haplotypes, primary and alternate, and both genome sizes are similar to the estimated value from GenomeScope 2.0 (Fig. 2A). The primary assembly consists of 289 scaffolds spanning 1.8 b with contig N50 of 37.3 Mb, scaffold N50 of 97 Mb, longest contig of 150.2 Mb and largest scaffold of 254.5 Mb. On the other hand, the alternate assembly consists of 224 scaffolds, spanning 1.7 Mb with contig N50 of 38.5 Mb, scaffold N50 of 107.2 Mb, longest contig 113 Mb, and largest scaffold of 198.8 Mb. Assembly statistics are reported in tabular form in Table 2, and graphical representation for the primary assembly in Fig. 2B.

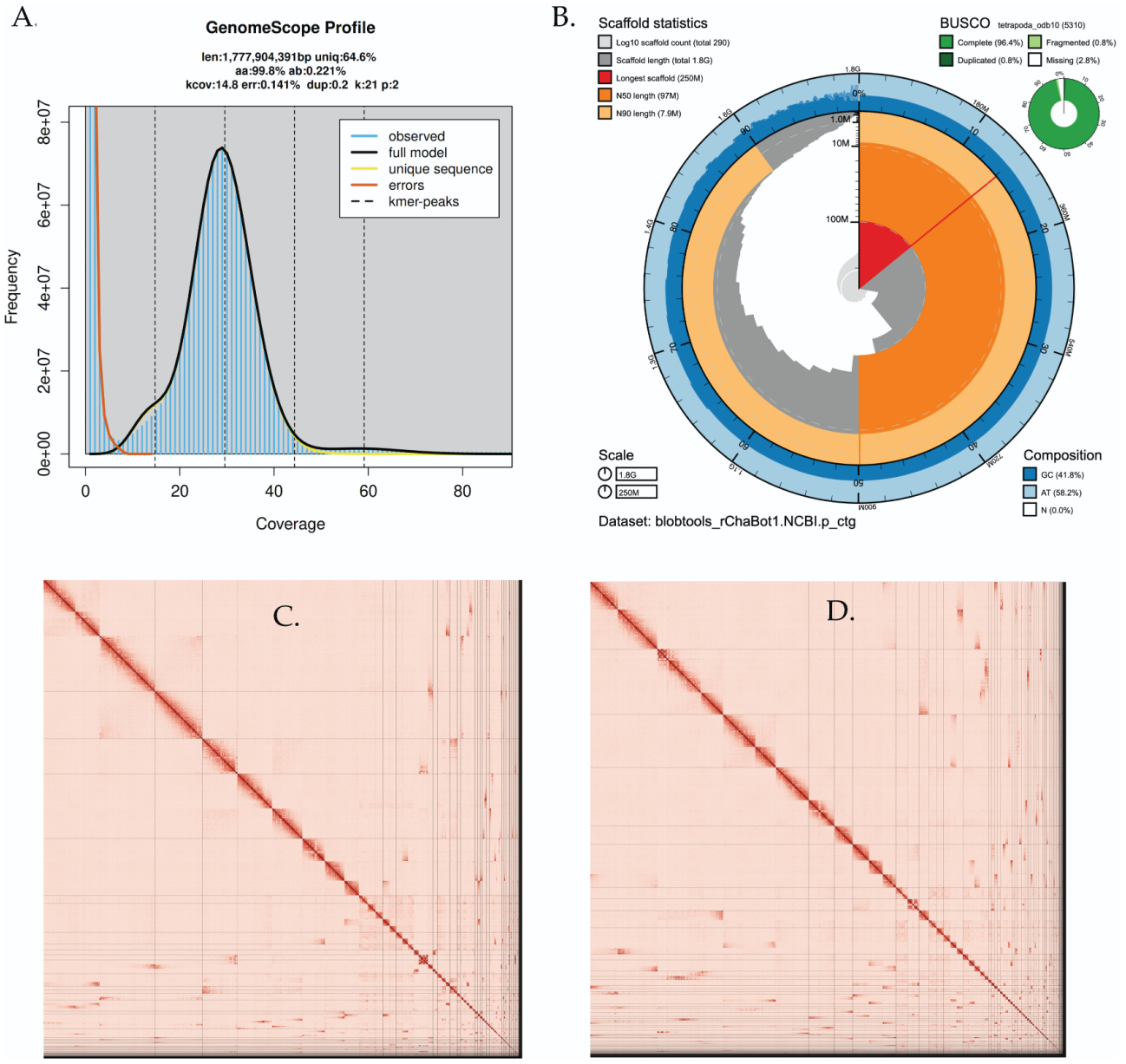


**Fig. 2.** Visual overview of genome assembly metrics. (A) K-mer spectra output generated from PacBio HiFi data without adapters using GenomeScope 2.0. The bimodal pattern observed corresponds to a diploid genome. K-mers covered at lower coverage but higher frequency correspond to differences between haplotypes, whereas the higher coverage but lower frequency k-mers corresponds to the similarities between haplotypes. (B) BlobToolKit Snail plot showing a graphical representation of the quality metrics presented in Table 2 for the *Charina bottae bottae* primary assembly (rActMar1). 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 versus light blue area around it shows mean, maximum and minimum GC versus AT content at 0.1% intervals (Challis et al. 2020). Hi-C contact maps for the primary (C) and alternate (D) genome assembly generated with PretextSnapshot. Hi-C contact maps translate proximity of genomic regions in 3D space to contiguous linear organization. Each cell in the contact map corresponds to sequencing data supporting the linkage (or join) between 2 of such regions. Scaffolds are separated by black lines and higher density corresponds to higher levels of fragmentation.

We identified a total of 1 misassembly on the alternate assembly and broke the corresponding join made by SALSA2, and closed a total of 4 gaps, per assembly. We further filtered out 2 contigs corresponding to mitochondrial contamination (1 per assembly). No further contigs were removed. The primary assembly has a BUSCO completeness score of 96.3% using the tetrapoda gene set, a per base quality (QV) of 60.4, a k-mer completeness of 97.2 and a frameshift indel QV of 46.6. The alternate assembly has a BUSCO completeness score of 91.2% using the tetrapoda gene set, a per base quality (QV) of 60.4, a k-mer completeness of 92.2 and a frameshift indel QV of 46.9. The Omni-C contact maps indicate that both assemblies are highly contiguous with some chromosome-length scaffolds (Fig. 2C and D). We have deposited scaffolds corresponding to both primary and alternate haplotype (see Table 2 and data availability for details).

**Table 2.** Sequencing and assembly statistics, and accession numbers.

| | | | | | | |
|---|---|---|---|---|---|---|
| BioProjects and vouchers | CCGP NCBI BioProject | | PRJNA720569 | | | |
| | Genera NCBI BioProject | | PRJNA766290 | | | |
| | Species NCBI BioProject | | PRJNA824832 | | | |
| | NCBI BioSample | | SAMN26367999 | | | |
| | Specimen identification | | HBS 135849 | | | |
| | NCBI Genome accessions | | Primary | Alternate | | |
| | Assembly accession | | JALMGJ000000000 | JALMGK000000000 | | |
| | Genome sequences | | GCA_023362775.1 | GCA_023362785.1 | | |
| Genome sequence | PacBio HiFi reads | Run | 1 PACBIO_SMRT (Sequel II) run: 3.6M spots, 54.2G bases, 35 Gb | | | |
| | | Accession | SRX15651215 | | | |
| | Omni-C Illumina reads | Run | 2 ILLUMINA (Illumina NovaSeq 6000) runs: 123.5M spots, 37.3G bases, 12.4 Gb | | | |
| | | Accession | SRX15651216, SRX15651217 | | | |
| Genome Assembly Quality Metrics | Assembly identifier (quality code[a]) | | rChaBot1(7.7.P7.Q60.C79) | | | |
| | HiFi read coverage[b] | | 30.96× | | | |
| | | | Primary | Alternate | | |
| | Number of contigs | | 374 | 308 | | |
| | Contig N50 | | 37,334,273 | 38,502,584 | | |
| | Contig NG50[b] | | 40,986,358 | 33,654,200 | | |
| | Longest contigs | | 150,241,728 | 113,059,615 | | |
| | Number of scaffolds | | 289 | 224 | | |
| | Scaffold N50 | | 97,015,800 | 107,200,052 | | |
| | Scaffold NG50[b] | | 97,015,800 | 107,200,052 | | |
| | Largest scaffold | | 254,579,594 | 198,876,488 | | |
| | Size of final assembly | | 1,804,944,895 | 1,703,974,979 | | |
| | Phased blocks NG50[b] | | 40,986,358 | 39,017,728 | | |
| | Gaps per Gbp (#Gaps) | | 47 (85) | 49 (84) | | |
| | Indel QV (frameshift) | | 47.64606828 | 48.12498448 | | |
| | Base pair QV | | 60.4101 | 60.443 | | |
| | | | Full assembly = 60.426 | | | |
| | K-mer completeness | | 97.2879 | 92.236 | | |
| | | | Full assembly = 99.2086 | | | |
| | BUSCO completeness (tetrapoda), *n* = 5,310 | | C | S | D | F | M |
| | | P[c] | 96.30% | 95.50% | 0.80% | 0.80% | 2.90% |
| | | A[c] | 91.20% | 90.40% | 0.80% | 1.00% | 7.80% |
| | Organelles | | 1 partial mitochondrial sequence | | | JALMGJ010000289.1 | |

[a]Assembly quality code *x*·*y*·*P*·*Q*·*C*, where *x* = log10[contig NG50]; *y* = log10[scaffold NG50]; *P* = log10[phased block NG50]; *Q* = Phred base accuracy QV (quality value); *C* = % genome represented by the first "*n*" scaffolds, following a known karyotype 2*n* = 36 for *C. bottae* (Gorman and Gress 1970). BUSCO scores. (C)omplete and (S)ingle; (C)omplete and (D)uplicated; (F)ragmented and (M)issing BUSCO genes. *n*, number of BUSCO genes in the set/database.
[b]Read coverage and NGx statistics have been calculated based on a genome size of 288 Mb.
[c]P(rimary) and (A)lternate assembly values.

We assembled a mitochondrial genome with MitoHiFi. The final mitochondrial genome size was 18,077 bp. The base composition of the final assembly version is A = 33.94%, C = 28.75%, G = 13.00%, T = 24.31%, and consists of 22 unique transfer RNAs and 13 protein coding genes.

## Discussion

There are currently reference genomes from 38 species representing 7 families and the *C. bottae bottae* genome is the first for the Boidae family. *Charina* is a member is the sister group to all caenophidian snakes that comprise the majority of global species diversity and is an important radiation of the snake tree of life (Gauthier et al. 2012; Simões and Pyron 2021). The closest relative of *Charina* with an assembled genome is *Python bivittatus* (family Pythonidae, Castoe et al. 2011, 2013). The *P. bivittatus* genome has a contig N50 length of 10.7 kb and an N50 scaffold length of 207.5 kb, whereas *C. bottae bottae* has a contig and scaffold N50 of 37.3 and 97 Mb. At 1.76 Gb, the *C. bottae bottae* genome is

about 20% larger than the *P. bivittatus* at 1.44Gb. These are the only 2 genomes from this diverse clade of snakes and will serve as an anchor for broader genomic comparisons across the phylogenetic breadth of macrostomatan (or the so-called "advanced") snakes.

The *C. bottae bottae* genome will be a valuable tool as conservation planning for the 4 geographic lineages within the *C. bottae bottae* complex continues to mature. Grismer et al. (2022) developed a RADSeq dataset of 19,711 loci that revealed multiple locally endemic clades within *C. bottae bottae* and *C. bottae umbratica*. This reference genome will be an invaluable tool for mapping these RAD loci, providing the research community the ability to better understand both physical linkage and paralog identities within this new dataset compared to the de novo RAD assembly currently available. Given the challenges facing cold-temperature and moisture-dependent species with increasing annual temperatures, droughts, and wildfires, our ability to clearly defined conservation units will be crucial to management over the next 70 yr (Abatzoglou et al. 2016; Westerling 2018; Williams et al. 2019; Grismer et al. 2022).

Phylogeographic studies on *Charina* indicate there is a disagreement between relationships inferred from mitochondrial and nuclear datasets (Rodríguez-Robles et al. 2001; Toshima 2011; Grismer et al. 2022). Grismer et al. (2022) recovered 4 reciprocally monophyletic nuDNA lineages restricted to coastal California, PGB, the Sierra Nevada Mountains, and southern California that shared a common ancestor 16.9 MYA. In contrasts Rodríguez-Robles et al. (2001) and Toshima (2011) used mtDNA and demonstrated that there are high levels of paraphyly among the PGB, the Sierra Nevada Mountains, and southern California lineages identified from RADSeq data (Grismer et al. 2022), with only populations from coastal California recovered as monophyletic. The *C. bottae bottae* mitochondrial and reference genomes described here, in combination with both published RADseq and future CCGP-generated whole genome resequencing data should facilitate ongoing and future studies aimed at disentangling the biogeographic and genetic process that have contributed to the current rubber boa lineages and their distribution, as well as genes responsible for their unique, cold-tolerant physiology.

## 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].

## 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. Partial support was provided by Illumina for Omni-C sequencing. 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.

## Data availability

Data generated for this study are available under NCBI BioProject PRJNA824832. Raw sequencing data for sample HBS135849 (NCBI BioSample SAMN26367999) are deposited in the NCBI Short Read Archive (SRA) under SRX15651215 for PacBio HiFi sequencing data, and SRX15651216, SRX15651217 for the Omni-C Illumina sequencing data. GenBank accessions for both primary and alternate assemblies are GCA_023362775.1 and GCA_023362785.1; and for genome sequences JALMGJ000000000 and JALMGK000000000. The GenBank organelle genome assembly for the mitochondrial genome is JALMGJ010000289.1. Assembly scripts and other data for the analyses presented can be found at the following GitHub repository: www.github.com/ccgproject/ccgp_assembly.

## References

Abatzoglou JT, Kolden CA, Balch JK, Bradley BA. Controls on interannual variability in lightning-caused fire activity in the western US. *Environ Res Lett*. 2016;4:1–11.

Abdennur N, Mirny LA. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics*. 2020;36:311–316.

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 Res*. 2020;20:892–905.

Blainville HD. Description de quelques especes de reptiles de la Californie. *Nouv Ann Mus Hist Nat (Paris)*. 1835;4:233–296.

Brehme CS, Clark DR, Rochester CJ, Fisher RN. Wildfires alter rodent community structure across four vegetation types in southern California, USA. *Fire Ecol*. 2011;7:81–98.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL, et al. BLAST+: architecture and applications. *BMC Bioinf*. 2009;10:1–9.

Castoe TA, de Koning AJ, Hall KT, Yokoyama KD, Gu W, Smith EN, Feschotte C, Uetz P, Ray DA, Dobry J, et al. 2011. Sequencing the genome of the Burmese python (Python molurus bivittatus) as a model for studying extreme adaptations in snakes. *Genome Biology*. 2011;12(406):1–8.

Castoe TA, Jason de Koning AP, Hall KT, Card DC, Schield DR, Fujita MK, Ruggieroa R, Degnerd JF, Daza JA, Gu W., et al. The Burmese python genome reveals the molecularbasis for extreme adaptation in snakes. *Proceed Nat Academy Sci*. 2013;110(51):20645–20650.

CDFG. *State and federally listed endangered and threatened animals of California*. Sacramento (CA): California Department of Fish and Game; 2005.

Challis R, Richards E, Rajan J, Cochrane G, Blaxter M. BlobToolKit—interactive quality assessment of genome assemblies. *G3*. 2020;10:1361–1374.

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods*. 2021;18:170–175.

Devitt T, Devitt S, Hollingsworth B, McGuire J, Moritz C. Montane refugia predict population genetic structure in the Large-blotched *Ensatina* salamander. *Mol Ecol*. 2013;22:1650–1665.

Gauthier JA, Kearney M, Maisano JA, Rieppel O, Behlke ABD. Assembling the squamate tree of life: perspectives from the phenotype and the fossil record. *Bulletin Peabody Museum Natural History*. 2012;53(1):3–308.

Ghurye J, Pop M, Koren S, Bickhart D, Chin CS. Scaffolding of long read assemblies using long range contact information. *BMC Genomics*. 2017;18:1–11.

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 Comput Biol*. 2019;8:e1007273.

Goloborodko A, Abdennur N, Venev S, Hbbrandao G. mirnylabS/pairtools: v0.2.0. 2018. doi:10.5281/zenodo.1490831.

Gorman GC, Gress F. Chromosome cytology of four boid snakes and a varanid lizard, with comments on the cytosystematics of primitive snakes. *Herpetologica*. 1970;26(3):308–317.

Grismer J, Scott P, Toffelmier E, Hinds B, Klabacka R, Stewart G, White V, Oaks J, Bradley Shaffer H. Genomic data reveal local endemism in Southern California Rubber Boas (Serpentes: Boidae, *Charina*) and the critical need for enhanced conservation actions. *Mol Phylogenet Evol*. 2022;174:1–17.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. 2013;29:1072–1075.

Kerpedjiev P, Abdennur N, Lekschas F, McCallum C, Dinkla K, Strobelt H, Luber JM, Ouellette SB, Azhir A, Kumar N, et al. HiGlass: web-based visual exploration and analysis of genome interaction maps. *Genome Biol*. 2018;19:1–12.

Klauber LM. The subspecies of the rubber boa, *Charina*. *Trans San Diego Soc Nat Hist*. 1943;10(7):83–90.

Korlach J, Gedman G, Kingan SB, Chin CS, Howard JT, Audet JN, 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.

Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, *arXiv*, arXiv:1303.3997, 2013, preprint: not peer reviewed. doi:10.48550/arXiv.1303.3997

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, *arXiv*, arXiv:2106.11799 [q-bio], 2021, preprint: not peer reviewed. doi:10.1093/molbev/msab199

Minnich RA, Franco-Vizcaíno E. Baja California's enduring Mediterranean vegetation: early accounts, human impacts, and conservation status. In: Cartron JLE, Ceballos G, editors. *Biodiversity, ecosystems, and conservation in Northern Mexico*. New York (NY): Oxford University Press; 2005. p. 370–386.

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;9:1–15.

Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun*. 2020;11:1–10.

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.

Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol*. 2020;21:1–27.

Rodríguez-Robles JA, Stewart GR, Pappenfuss TJ. Mitochondrial DNA-based phylogeography of North American rubber boas, *Charina bottae* (Serpentes: Boidae). *Mol Phylogenet Evol*. 2001;18:227–237.

Seppey M, Manni M, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness. *Methods Mol Biol*. 2019;1962:227–245.

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.

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: 1–7.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 2015;31:3210–3212.

Simões TR, Pyron RA. The squamate tree of life. *Bulletin Museum of Comparative Zool*. 2021;163(2):47–95.

Stebbins RC. *A field guide to western reptiles and amphibians*. 3rd ed. New York (NY): Houghton Mifflin Company; 2003.

Stewart GR. *Charina b. bottae*. *Catalogue of American Amphibians and Reptiles*. St. Louis (MO): Society for the Study of Amphibians and Reptiles; 1977.

Syphard AD, Butsic V, Bar-Massada A, Keeley JE, Tracey JA, Fisher RN. Setting priorities for private land conservation in fire-prone landscapes: are fire risk reduction and biodiversity conservation competing or compatible objectives? *Ecol Soc*. 2016;3:1–11.

Todd BD, Jenkinson TS, Escalona M, Beraut E, Nguyen O, Sahasrabudhe R, Scott PA, Toffelmier E, Wang IJ, Shaffer HB. Reference genome of the northwestern pond turtle, *Actinemys marmorata*. *J Hered*. 2022;113:624–631.

Tracey JA, Rochester CJ, Hathaway SA, Preston KL, Syphard AD, Vandergast AG, Diffendorfer JE, Franklin J, MacKenzie JB, Oberbauer TA. et al. Prioritizing conserved areas threatened by wildfire and fragmentation for monitoring and management. *PLoS One*. 2018;19:e0200203.

Van Denburgh J. Description of a new subspecies of boa (*Charina bottae utahensis*) from Utah. *Proc Calif Acad Sci*. 1920;10(3):31–32.

Vandergast AG, Bohonak AJ, Hathaway SA, Boys J, Fisher RN. Are hotspots of evolutionary potential adequately protected in southern California? *Biol Conserv*. 2008;141:1648–1664.

Westerling AL. Wildfire simulations for California's fourth climate change assessment: projecting changes in extreme wildfire events with a warming climate. California's Fourth Climate Change Assessment, Rep. CCCA4-CEC-2018-014. California Energy Commission; 2018. University of California, Merced.

Williams AP, Abatzoglou JT, Gershunov A, Guzman-Morales J, Bishop DA, Balch JK, Lettenmaier DP. Observed impacts of anthropogenic climate change on wildfire in California. *Earth's Future*. 2019;7:892–910.

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(6): 632–640.