

**Assembly of the largest squamate reference genome to date: the western fence lizard,
*Sceloporus occidentalis***

Anusha P. Bishop^{1,2}, Erin P. Westeen^{1,2}, Michael L. Yuan³, Merly Escalona⁴, Eric Beraut⁵, Colin Fairbairn⁵, Mohan P. A. Marimuthu⁶, Oanh Nguyen⁶, Noravit Chumchim⁶, Erin Toffelmier^{7,8}, Robert N. Fisher⁹, H. Bradley Shaffer^{7,8}, and Ian J. Wang^{1,2}

Author affiliation(s):

¹ Department of Environmental Science, Policy, and Management, University of California, Berkeley, CA 94720, USA

² Museum of Vertebrate Zoology, University of California, Berkeley, CA 94720, USA

³ Center for Population Biology, University of California, Davis, CA 95616, USA

⁴ Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA 95064, USA.

⁵ Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, Santa Cruz, CA 95064, USA

⁶ DNA Technologies and Expression Analysis Core Laboratory, Genome Center, University of California, Davis, CA 95616, USA

⁷ Department of Ecology and Evolutionary Biology, University of California, Los Angeles, CA 90095, USA

⁸ La Kretz Center for California Conservation Science, Institute of the Environment and Sustainability, University of California, Los Angeles, CA 90095, USA

⁹ U.S. Geological Survey Western Ecological Research Center, San Diego, CA 92101, USA

Corresponding author email: ianwang@berkeley.edu

Running title: Western fence lizard genome

Abstract

Spiny lizards (genus *Sceloporus*) have long served as important systems for studies of behavior, thermal physiology, dietary ecology, vector biology, speciation, and biogeography. The western fence lizard, *Sceloporus occidentalis*, is found across most of the major biogeographical regions in the western United States and northern Baja California, Mexico, inhabiting a wide range of habitats, from grassland to chaparral to open woodlands. As small ectotherms, *Sceloporus* lizards are particularly vulnerable to climate change, and *S. occidentalis* has also become an important system for studying the impacts of land use change and urbanization on small vertebrates. Here, we report a new reference genome assembly for *S. occidentalis*, as part of the California Conservation Genomics Project (CCGP). Consistent with the reference genomics 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 a total of 608 scaffolds spanning 2,856 Mb, has a contig N50 of 18.9 Mb, a scaffold N50 of 98.4 Mb, and BUSCO completeness score of 98.1% based on the tetrapod gene set. This reference genome will be valuable for understanding ecological and evolutionary dynamics in *S. occidentalis*, the species status of the California endemic island fence lizard (*S. becki*), and the spectacular radiation of *Sceloporus* lizards.

Keywords

California Conservation Genomics Project, CCGP, reptile, Iguania, Phrynosomatidae, de novo genome assembly, Squamata

The archetypal western fence lizard exhibits light brown to dark gray dorsal coloration with dark brown to black chevrons in two longitudinal lines down the back, yellow to orange coloration on the posterior of all limbs, and bright blue to blue-green ventral patches, particularly in adult males, that make them easily recognizable and have given them the colloquial name “blue-bellies.” However, they display phenotypic variation across their range, including distinct high elevation (Leache et al. 2010) and urban (Putman et al. 2019) phenotypes, revealing the

1
2
3 46 tight coupling between local environmental conditions and external traits. Despite their large
4
5 47 range and ability to persist across many habitat types, as small ectotherms these lizards are
6
7 48 sensitive to changes in their thermal environment, and spiny lizards (genus *Sceloporus*) are
8
9 49 projected to experience significant declines due to climate change (Sinervo et al. 2010).
10
11 50 *Sceloporus occidentalis* populations are also frequently affected by urbanization, land use
12
13 51 conversion, and habitat fragmentation (Delaney et al. 2010, Delaney et al. 2021), making the
14
15 52 development of genomic resources in this system valuable for studies of organismal responses to
16
17 53 landscape alteration and climate change, for which *S. occidentalis* could serve as a model
18
19 54 system.

20
21
22 55 Here, we present the first assembled reference genome for *Sceloporus occidentalis*,
23
24 56 produced as part of the California Conservation Genomics Project (CCGP; Shaffer et al. 2022,
25
26 57 Toffelmier et al. 2022). The *S. occidentalis* genome assembly joins 42 other lizard reference
27
28 58 genomes, including four others for Phrynosomatidae, a family with roughly 140 species. This
29
30 59 genome assembly will contribute to our understanding of the ecology and evolution of this
31
32 60 widespread, ecologically and scientifically important, charismatic species. Because *S.*
33
34 61 *occidentalis* is an abundant member of most communities in California, this new resource will
35
36 62 also contribute to several ongoing and future aspects of California habitat and biodiversity
37
38 63 conservation (Amburgey et al. 2021, Fiedler et al. 2022).
39
40
41
42
43
44
45
46

47 **Methods**

48 ***Biological Materials***

49
50
51 67 We collected an adult male *S. occidentalis* (field number: IW3139) from Yosemite
52
53 68 National Park, Mariposa County, California (37.72642, -119.38959), in August 2020. The
54
55
56
57
58
59
60

specimen was collected under applicable state (SC-8436) and national park (YOSE-2020-SCI-0058) scientific research permits following a protocol approved by the UC Berkeley Animal Care and Use Committee (AUP-2016-02-8453-2). We removed liver tissue, immediately flash froze it in liquid nitrogen, and stored it at -80 °C until extraction of genomic DNA.

DNA extraction

High molecular weight (HMW) genomic DNA (gDNA) was extracted from 30 mg of liver tissue using the Nanobind Tissue Big DNA kit following the manufacturer's instructions (Pacific BioSciences - PacBio, Menlo Park, CA). DNA purity was estimated using absorbance ratios ($260/280 = 1.85$ and $260/230 = 2.40$) on a NanoDrop ND-1000 spectrophotometer. The final DNA yield (298 ng/μl; 37 μg) was quantified using a 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 indicated that 63% of the fragments were 100 Kb or longer.

HiFi library preparation and sequencing

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 kb – 20 kb. The sheared gDNA was concentrated using 0.45X of AMPure PB beads (PacBio, Cat. #100-265-900) for the removal of single-strand overhangs at 37 °C for 15 minutes, followed by further enzymatic steps of DNA damage repair at 37 °C for 30 minutes, end repair and A-tailing at 20 °C for 10 minutes and 65C for 30 minutes, ligation of overhang adapter v3 at 20 °C for 60 minutes and 65 °C for

10 minutes to inactivate the ligase, then nuclease treated at 37 °C for 1 hour. The SMRTbell library was purified and concentrated with 0.45X 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 7-9 kb. The 15 – 20 kb average HiFi SMRTbell library was sequenced at the University of California Davis DNA Technologies Core (Davis, CA) using five 8M SMRT cells, Sequel II sequencing chemistry 2.0, and 30-hour 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, after which chromatin was fixed in place in the nucleus. The suspended chromatin solution was then passed through 100 µm 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, Ipswich, MA) with an Illumina compatible y-adaptor. Biotin-containing fragments were then captured using streptavidin beads. The post-capture 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 x 150 bp read pairs per GB of genome size.

Nuclear genome assembly

We assembled the *S. occidentalis* genome following the CCGP assembly pipeline Version 4, which uses PacBio HiFi reads and Omni-C data to produce high quality and highly contiguous genome assemblies while minimizing manual curation (outlined on 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) with the filtered PacBio HiFi reads and the Omni-C dataset. We tagged output haplotype 1 as the primary assembly, and output haplotype 2 as the alternate assembly. We identified sequences corresponding to haplotypic duplocations, contig overlaps and repeats on the primary assembly with purge_dups (Guan et al 2020) and transferred them to the alternate assembly. We scaffolded both assemblies using the Omni-C data with SALSA (Ghurye et al. 2017; Ghurye et al. 2019).

We generated Omni-C contact maps by aligning the Omni-C data against both assemblies with BWA-MEM (Li 2013) and identified ligation junctions and generated Omni-C pairs using pairtools (Golobordko et al. 2018). We generated a multi-resolution Omni-C matrix with cooler (Abdennur an Mirny 2020) and balanced it with hicExplorer (Ramírez et al. 2018). We used HiGlass [Version 2.1.11] (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 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 NCBI contamination screening.

Mitochondrial genome assembly

We assembled the mitochondrial genome of *S. occidentalis* from the PacBio HiFi reads using the reference-guided pipeline MitoHiFi (<https://github.com/marcelauliano/MitoHiFi>; Allio et al. 2020). An existing mitochondrial sequence for *S. occidentalis* (NCBI:AB079242.1) 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.

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

1
2
3 161 et al. 2021) with the Tetrapoda ortholog database (tetrapoda_odb10), which contains 5,310
4
5 162 genes. Assessments of base-level accuracy (QV) and k-mer completeness were performed using
6
7 163 the previously generated meryl database and merqury (Rhie et al. 2020). We further estimated
8
9
10 164 genome assembly accuracy via BUSCO gene set frameshift analysis using the pipeline described
11
12 165 in Korlach et al. (2017).

13
14 166 Measurements of the size of the phased blocks is based on the size of the contigs
15
16
17 167 generated by HiFiasm in HiC mode. We followed the quality metric nomenclature established by
18
19 168 Rhie et al. (2021), with the genome quality code x.y.P.Q.C, where, x = log10[contig NG50]; y =
20
21 169 log10[scaffold NG50]; P = log10 [phased block NG50]; Q = Phred base accuracy QV (quality
22
23 170 value); and C = % genome represented by the first ‘n’ scaffolds, following a known karyotype of
24
25 171 2n =22 for *S. occidentalis* (Jackson & Hunsacker 1970). Quality metrics for the notation were
26
27 172 calculated on the primary assembly.
28
29
30

31 173
32
33 174 **Results**

34
35 175 The Omni-C and PacBio HiFi sequencing libraries generated 290 million read pairs and
36
37 176 5.3 million reads respectively. The latter yielded 29.94-fold coverage (N50 read length 16,535
38
39 177 bp; minimum read length 45 bp; mean read length 16,088 bp; maximum read length of 63,768
40
41 178 bp) based on the Genomescope 2.0 genome size estimation of 2.8 Gb. Based on PacBio HiFi
42
43 179 reads, we estimated 0.168 % sequencing error rate and 0.521% nucleotide heterozygosity rate.
44
45 180 The k-mer spectrum based on PacBio HiFi reads show a bimodal distribution with a major peak
46
47 181 at ~29-fold coverage (Figure 2A), which supports that of a low heterozygosity profile.
48
49
50

51 182 The final assembly (rSceOcc1) consists of two pseudo haplotypes, primary and alternate.
52
53 183 Genome size for primary assembly is similar to the estimated value from Genomescope2.0
54
55
56
57
58
59
60

(Figure 2A), but the alternate is ~400Mb larger. The primary assembly consists of 608 scaffolds spanning 2.8 Gb with contig N50 of 18.9 Mb, scaffold N50 of 98.4 Mb, longest contig of 124.1 Mb and largest scaffold of 320.8 Mb. The alternate assembly consists of 1,822 scaffolds, spanning 3.2 Gb with contig N50 of 17.6 Mb, scaffold N50 of 38.7 Mb, largest contig 115.5 Mb and largest scaffold of 309.6 Mb. Detailed assembly statistics are reported in tabular form in Table 2 and graphical representation for the primary assembly in Figure 2B (See Supplementary Figure 1 for the alternate assembly). The primary assembly has a BUSCO completeness score of 98.1% using the Tetrapoda gene set, a per- base quality (QV) of 61.68, a k-mer completeness of 92.89 and a frameshift indel QV of 48.71, while the alternate assembly has a BUSCO completeness score of 98.3% using the same gene set, a per- base quality (QV) of 61.39, a k-mer completeness of 93.62 and a frameshift indel QV of 49.12.

We identified 2 misassemblies on the alternate assembly and broke the corresponding joins made by SALSA. We were able to close a total of 9 gaps, 7 on the primary assembly and 2 on the alternate. Finally, we filtered out 34 contigs, 2 from the primary and 32 from the alternate assembly, corresponding to mitochondrial contamination. No further contigs were removed. We have deposited both assemblies on NCBI (see Table 2 and Data Availability for details).

We assembled a mitochondrial genome using MitoHiFi. The final mitochondrial genome assembly size is 14,426 bp. The final assembly version has a base composition of A=39.95%, C=17.11%, G=10.38%, T=32.56% and consists of 20 unique transfer RNAs and 12 protein coding genes.

Discussion

There are currently 89 other squamate genomes, including 42 lizards and 47 snakes, available on NCBI, ranging in size from 1.13 Gb to 2.74 Gb (mean = 1.64 Gb), making the *Sceloporus occidentalis* genome the largest squamate reference genome assembled so far. At 2.86 Gb, the *S. occidentalis* genome is 4.2% larger than the next largest, the Japanese gecko, *Gekko gekko*. The 89 other squamate reference genomes have an average contig N50 of 14.5 Mb (ranging from 1.1 Kb to 127.8 Mb), placing the contig N50 of 19.0 Mb for our *S. occidentalis* assembly in the upper end of this range, below only 17 other reference genomes.

Three of these existing 89 reference genomes are also phrynosomatid lizards, including two other *Sceloporus* species (*S. tristichus* and *S. undulatus*). The genome assembly sizes of these other species range from 1.82 Gb to 1.91 Gb (mean = 1.87 Gb), so the *S. occidentalis* genome is considerably larger than those of its congeners. Both of these species and *S. occidentalis* are members of the same phylogenetic clade within *Sceloporus*, the *undulatus* group, which has a crown age of 15-20 million years (Leache et al. 2016). Until recently *S. tristichus* and *S. undulatus* were considered conspecific lineages, and so their very similar genome assembly sizes are not surprising. However, the differences in putative genome sizes between these two close relatives and *S. occidentalis* suggest either substantial genome size evolution within the genus or the under- or over-estimation of genome size for one or more of these species. The C value for *S. occidentalis* is 2.36 Gb, based on Feulgen densitometry (Olmo 1981), suggesting a genome size larger than the 1.82 Gb *S. tristichus* (Bedoya and Leache 2021) and 1.91 Gb *S. undulatus* (Westfall et al. 2021) genome assemblies. The description of the *S. undulatus* genome indicates that the assembly may be missing some data or has repeat regions condensed (Westfall et al. 2021). It is also possible that our current assembly overestimated the

size of the *S. occidentalis* genome, although this seems unlikely given the high BUSCO completeness (98.4%, see Table 2) and the relatively large estimate based on Feulgen densitometry (Olmo 1981).

The *S. occidentalis* reference genome provides a valuable resource for studying the conservation, ecology, and evolutionary dynamics of this species and other spiny lizards. The genus *Sceloporus* constitutes a rapid radiation containing more than 100 species, and the *Sceloporus* phylogeny as currently understood is characterized by high levels of gene tree discordance (Leache et al. 2016). This reference genome can help to further resolve the species relationships in this speciose clade of lizards. It will also contribute to understanding patterns of speciation in this group, identifying regions of the genome involved in adaptation to environmental variation and rapid speciation, and examining the evolutionary dynamics in the complex biogeographic history of *S. occidentalis*.

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], with additional support from a National Science Foundation CAREER grant [DEB-1845682] to IJW.

Acknowledgements

PacBio Sequel II library prep and sequencing was 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, and we thank Andrew Crawford for providing helpful comments on the manuscript. We also thank the California Department of Fish and Wildlife (SC-8436) and Yosemite National Park (YOSE-2020-SCI-0058) for granting scientific research permits. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Data Availability

Data generated for this study are available under NCBI BioProject PRJNA720569. Raw sequencing data for sample IW3139 (NCBI BioSample SAMN27480378) are deposited in the NCBI Short Read Archive (SRA) under SRX15651255, SRX15651256, and SRX15651257. 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, L. A. (2020). Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics* , 36(1), 311–316. doi: 10.1093/bioinformatics/btz540

Adolph, S. C. 1990. Influence of behavioral thermoregulation on microhabitat use by two *Sceloporus* lizards. *Ecology* 71:315–327.

- 274 Allio, R., Schomaker-Bastos, A., Romiguier, J., Prosdocimi, F., Nabholz, B., & Delsuc, F.
275 (2020). MitoFinder: Efficient automated large-scale extraction of mitogenomic data in
276 target enrichment phylogenomics. *Molecular Ecology Resources*, 20(4), 892–905. doi:
277 10.1111/1755-0998.13160
- 278 Amburgey, S.M., Miller, D.A.W., Rochester, C.J., Delaney, K.S., Riley, S.P.D., Brehme, C.S.,
279 Hathaway, S.A., and R.N. Fisher. 2021. The influence of species life history and
280 distribution characteristics on species responses to habitat fragmentation in an urban
281 landscape. *Journal of Animal Ecology* 90:685-697.
- 282 Bedoya, A. M., and A. D. Leaché. 2021. Characterization of a pericentric inversion in plateau
283 fence lizards (*Sceloporus tristichus*): evidence from chromosome-scale genomes. *G3*
284 *Genes|Genomes|Genetics* 11:jkab036.
- 285 Bouzid, N. M., J. W. Archie, R. A. Anderson, J. A. Grummer, and A. D. Leaché. 2022. Evidence
286 for ephemeral ring species formation during the diversification history of western fence
287 lizards (*Sceloporus occidentalis*). *Molecular Ecology* 31:620–631.
- 288 Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL.
289 2009. BLAST+: architecture and applications. *BMC Bioinformatics*. 10:421.
- 290 Challis, R., Richards, E., Rajan, J., Cochrane, G., & Blaxter, M. (2020) BlobToolKit –
291 Interactive quality assessment of genome assemblies. *G3 Genes|Genomes|Genetics*.
292 10:1361-1374. doi.org/10.1534/g3.119.400908
- 293 Cheng H, Jarvis ED, Fedrigo O, Koepfli K-P, Urban L, Gemmell NJ, et al. Robust haplotype-
294 resolved assembly of diploid individuals without parental data. *arXiv [q-bio.GN]*. 2021.
295 Available: <http://arxiv.org/abs/2109.04785>
- 296 Delaney, K. S., S. P. D. Riley, and R. N. Fisher. 2010. A rapid, strong, and convergent genetic

1
2
3 297 response to urban habitat fragmentation in four divergent and widespread vertebrates.
4
5 298 PLOS ONE 5:e12767.
6
7
8 299 Delaney, K.S., Busteed, G., Fisher, R.N., and S.P.D. Riley. 2021. Reptile and amphibian
9
10 300 diversity and abundance in an urban landscape: Impacts of fragmentation and the
11
12 301 conservation value of small patches. Ichthyology & Herpetology 109:424-435.
13
14 302 Fiedler, P. L., B. Erickson, M. Esgro, M. Gold, J. M. Hull, J. Norris, B. Shapiro, M. Westphal,
15
16 303 E. Toffelmier, and H. B. Shaffer. 2022. Seizing the Moment: The opportunity and
17
18 304 relevance of the California Conservation Genomics Project to state and federal
19
20 305 conservation policy. Journal of Heredity 113:589-596.
21
22
23 306 Ghurye, J., Pop, M., Koren, S., Bickhart, D., & Chin, C.-S. (2017). Scaffolding of long read
24
25 307 assemblies using long range contact information. BMC Genomics, 18(1), 527. doi:
26
27 308 10.1186/s12864-017-3879-z
28
29
30 309 Ghurye, J., Rhie, A., Walenz, B. P., Schmitt, A., Selvaraj, S., Pop, M., ... Koren, S. (2018).
31
32 310 Integrating Hi-C links with assembly graphs for chromosome-scale assembly (p.
33
34 311 e1007273). doi: 10.1371/journal.pcbi.1007273
35
36
37 312 Goldberg, S. R. 1974. Reproduction in mountain and lowland populations of the lizard
38
39 313 *Sceloporus occidentalis*. Copeia 1974:176–182.
40
41
42 314 Goloborodko, A., Abdennur, N., Venev, S., hbbrandao, & gfudenberg. (2018).
43
44 315 mirnylab/pairtools: v0.2.0. doi: 10.5281/zenodo.1490831
45
46
47 316 Guan, D., McCarthy, S. A., Wood, J., Howe, K., Wang, Y., & Durbin, R. (2020). Identifying and
48
49 317 removing haplotypic duplication in primary genome assemblies. Bioinformatics , 36(9),
50
51 318 2896–2898. doi: 10.1093/bioinformatics/btaa025
52
53
54 319 Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUASt: quality assessment tool for
55
56
57
58
59
60

- genome assemblies. *Bioinformatics*, 29(8), 1072–1075. doi: 10.1093/bioinformatics/btt086
- Jackson, L. & Hunsaker, D. (1970) Chromosome morphology of sceloporine lizards (*Sceloporus occidentalis* and *S. graciosus*). *Experientia*, 26, 198–199.
- Kerpedjiev, P., Abdennur, N., Lekschas, F., McCallum, C., Dinkla, K., Strobelt, H., ... Gehlenborg, N. (2018). HiGlass: web-based visual exploration and analysis of genome interaction maps. *Genome Biology*, 19(1), 125. doi: 10.1186/s13059-018-1486-1
- Korlach, J., Gedman, G., Kingan, S. B., Chin, C.-S., Howard, J. T., Audet, J.-N., ... Jarvis, E. D. (2017). De novo PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. *GigaScience*, 6(10), 1–16. doi: 10.1093/gigascience/gix085
- Leaché, A. D., B. L. Banbury, C. W. Linkem, and A. N.-M. de Oca. 2016. Phylogenomics of a rapid radiation: is chromosomal evolution linked to increased diversification in North American spiny lizards (genus *Sceloporus*)? *BMC Evolutionary Biology* 16:63.
- Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*. Retrieved from <http://arxiv.org/abs/1303.3997>
- Liu, Y., Zhou, Q., Wang, Y. et al. Gekko japonicus genome reveals evolution of adhesive toe pads and tail regeneration. *Nat Commun* 6, 10033 (2015).
- Manni M., Berkeley M.R., Seppey M., Simao F.A., Zdobnov E.M. 2021. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *arXiv:2106.11799 [q-bio]* [Internet]. Available from: <http://arxiv.org/abs/2106.11799>
- Marcellini, D., and J. P. Mackey. 1970. Habitat preferences of the lizards, *Sceloporus*

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

occidentalis and *S. graciosus* (Lacertilia, Iguanidae). *Herpetologica* 26:51–56.

Olmo, E. (1981). Evolution of genome size and DNA base composition in reptiles. *Genetica* 57: 39-50

Pflug, J. M., V. R. Holmes, C. Burrus, J. S. Johnston, D. R. Maddison. 2020. Measuring genome sizes using read-depth, k-mers, and flow cytometry: methodological comparisons in beetles (Coleoptera). *G3 Genes|Genomes|Genetics*, 20: 3047-3060. <https://doi.org/10.1534/g3.120.401028>

Putman, B. J., M. Gasca, D. T. Blumstein, and G. B. Pauly. 2019. Downsizing for downtown: limb lengths, toe lengths, and scale counts decrease with urbanization in western fence lizards (*Sceloporus occidentalis*). *Urban Ecosyst* 22:1071–1081.

Ramírez, F., Bhardwaj, V., Arrigoni, L., Lam, K. C., Grüning, B. A., Villaveces, J., ... Manke, T. (2018). High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nature Communications*, Vol. 9. doi: 10.1038/s41467-017-02525-w

Ranallo-Benavidez, T. R., Jaron, K. S., & Schatz, M. C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications*, 11(1), 1432. doi: 10.1038/s41467-020-14998-3

Rhie, A., Walenz, B. P., Koren, S., & Phillippy, A. M. (2020). Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology*, 21(1), 245. doi: 10.1186/s13059-020-02134-9

Rhie A., McCarthy S. A., Fedrigo O., Damas J., Formenti G., Koren S., Uliano-Silva M., Chow W., Fungtammasan A., Kim J., et al. (2021). Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, 592:737–746. doi: 10.1038/s41586-021-03451-0

- Roll, U., A. Feldman, M. Novosolov, A. Allison, A. M. Bauer, R. Bernard, M. Böhm, F. Castro-Herrera, L. Chirio, B. Collen, G. R. Colli, L. Dabool, I. Das, T. M. Doan, L. L. Grismer, M. Hoogmoed, Y. Itescu, F. Kraus, M. LeBreton, A. Lewin, M. Martins, E. Maza, D. Meirte, Z. T. Nagy, C. de C. Nogueira, O. S. G. Pauwels, D. Pincheira-Donoso, G. D. Powney, R. Sindaco, O. J. S. Tallowin, O. Torres-Carvajal, J.-F. Trape, E. Vidan, P. Uetz, P. Wagner, Y. Wang, C. D. L. Orme, R. Grenyer, and S. Meiri. 2017. The global distribution of tetrapods reveals a need for targeted reptile conservation. *Nat Ecol Evol* 1:1677–1682.
- Rochester, C.J., C.S. Brehme, D.R. Clark, D. Stokes, S.A. Hathaway, and R.N. Fisher. 2010. Reptile and amphibian responses to large-scale wildfires in southern California. *Journal of Herpetology* 44:333-351.
- Rose, B. R. 1976. Habitat and prey selection of *Sceloporus occidentalis* and *Sceloporus graciosus*. *Ecology* 57:531–541.
- Salerno, P. E., Chan, L. M., Pauly, G. B., Funk, W. C., and Robertson, J. M. 2023. Near-shore island lizard fauna shaped by a combination of human-mediated and natural dispersal. *Journal of Biogeography* 50: 116–129.
- Shaffer, H. B., E. Toffelmier, R. B. Corbett-Detig, M. Escalona, B. Erickson, P. Fiedler, M. Gold, R. J. Harrigan, S. Hodges, T. K. Luckau, C. Miller, D. R. Oliveira, K. E. Shaffer, B. Shapiro, V. L. Sork, and I. J. Wang. 2022. Landscape genomics to enable conservation actions: The California Conservation Genomics Project. *Journal of Heredity* 113:577-588.
- Sim, S. (2021). sheinasim/HiFiAdapterFilt: First release. doi: 10.5281/zenodo.4716418

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Sinervo, B., and J. B. Losos. 1991. Walking the tight rope: arboreal sprint performance among
Sceloporus occidentalis lizard populations. Ecology 72:1225–1233.

Stebbins, R. C., and S. M. McGinnis. 2012. Field Guide to Amphibians and Reptiles of
California. University of California Press, Berkeley, CA

Toffelmier, E., J. Beninde, and H. B. Shaffer. 2022. The phylogeny of California, and how it
informs setting multi-species conservation priorities. Journal of Heredity 113:597-603.
doi.org/10.1093/jhered/esac045

Westfall, A. K., R. S. Telemeco, M. B. Grizante, D. S. Waits, A. D. Clark, D. Y. Simpson, R. L.
Klabacka, A. P. Sullivan, G. H. Perry, M. W. Sears, C. L. Cox, R. M. Cox, M. E. Gifford,
H. B. John-Alder, T. Langkilde, M. J. Angilletta, A. D. Leaché, M. Tollis, K. Kusumi,
and T. S. Schwartz. 2021. A chromosome-level genome assembly for the Eastern Fence
Lizard (*Sceloporus undulatus*), a reptile model for physiological and evolutionary
ecology. GigaScience 10:1–17.

Wishingrad, V., & Thomson, R. C. (2023). Biogeographic inferences across spatial and
evolutionary scales. Molecular Ecology 32:2055-2070.

Table 1: Assembly pipeline and software used. 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
<i>De novo assembly (contig)</i>	HiFiasm (Hi-C Mode, -primary, output p_ctg.hap1, p_ctg.hap2)	0.16.1-r375
Remove low-coverage, duplicated contigs	purge_dups	1.2.6
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
	PretextView	0.1.4
	PretextView	0.1.5
	PretextViewSnapshot	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
	Merquery	2020-01-29
Contamination screening		
Local alignment tool	BLAST+	2.1
General contamination screening	BlobToolKit	2.3.3

Table 2: Sequencing and assembly statistics and accession numbers.

Bio Projects & Vouchers	CCGP NCBI BioProject	PRJNA720569
	Genus NCBI BioProject	PRJNA765857
	Species NCBI BioProject	PRJNA777217
	NCBI Bio-sample	SAMN27480378
	Specimen identification number	IW3139
Genome Sequence	PacBio HiFi long read runs	1 PacBio SMRT Sequel II run: 5.7M spots, 85.9G bp, 57.2Gb
	OmniC Illumina sequencing	1 Illumina NovaSeq 6000 run: 290M spots, 87.5G bp, 28Gb
	PacBio HiFi NCBI SRA Accession	SRX15651255
	OmniC Illumina NCBI SRA Accession	SRX15651256, SRX15651257
Genome Assembly Primary (Alternate)	Assembly identifier	rSceOcc1
	HiFi Read coverage	37.14X
	Number of contigs (primary / alternate)	659 / 1822
	Contig N50 (bp)	18,989,278 / 17, 628, 953
	Contig NG50 [§] (bp)	18,989,278 / 20,278,101
	Longest Contigs (primary/ alternate)	124,125,603 / 115,579,027
	Number of scaffolds (primary/ alternate)	608 / 1,771
	Scaffold N50 (bp)	98,418,489 / 38,771,511
	Scaffold NG50 [§]	98,418,489 / 88,220,525
	Size of final assembly (bp)	2,856,356,971 / 3,186,658,811
	Gaps per Gbp (# Gaps)	18 (51) / 16 (51)
Assembly Quality ‡	NCBI Genome Assembly Accession	GCA_000XXXXXX.1
	Assembly quality identifier*	7.7.P7.Q61.C57
	Base pair QV (Merqury)	P: Q 61.6835, A: Q 61.4466
	Indel QV (Frame shift analysis)	P: Q 48.7129, A: Q 49.1278
	k-mer completeness	P: 92.8931%, A: 93.6217%
	BUSCO completeness	
	Primary (C:S:D:F:M)	98.10% : 32.90% : 65.20% : 0.80% : 1.10%
	Alternate (C:S:D:F:M)	98.30% : 31.90% : 66.40% : 0.70% : 1.00%
	Organelles	1 complete mitochondrial sequence CM041364.1

* Assembly quality code x.y.P.Q.C derived notation, from (Rhie et al. 2021). 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 for SPECIES of 2n=22. Quality code for all the assembly denoted by primary assembly (rSceOcc1.0.p)

§ Read coverage and NGx statistics have been calculated based on the estimated genome size of 2.85 Gb

‡ (P)primary and (A)lternate assembly values.

Figure Legends

Figure 1. Photo of the western fence lizard (*Sceloporus occidentalis*) used to generate the reference genome (A) and a range map (B) for *S. occidentalis* based on data from the Global Assessment of Reptile Distributions (Roll et al. 2017) with the location for the genome specimen indicated by the yellow circle. Western fence lizards occupy habitat ranging from montane mixed conifer forest (C) to open woodland and grassland (D) to coastal scrub (E). Note the basking *S. occidentalis* on the log in the foreground of panel E.

Figure 2. Visual overview of genome assembly metrics. A) K-mer spectrum output generated from PacBio HiFi data without adapters using GenomeScope2.0. B) BlobToolKit Snail plot showing a graphical representation of the quality metrics presented in Table 2 for the *S. occidentalis* primary assembly (rSceOcc1). 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-D) 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 3-D space to contiguous linear organization. Each cell in the contact map corresponds to sequencing data supporting the linkage (or join) between two such regions. Scaffolds are separated by black lines, and higher density corresponds to higher levels of fragmentation.

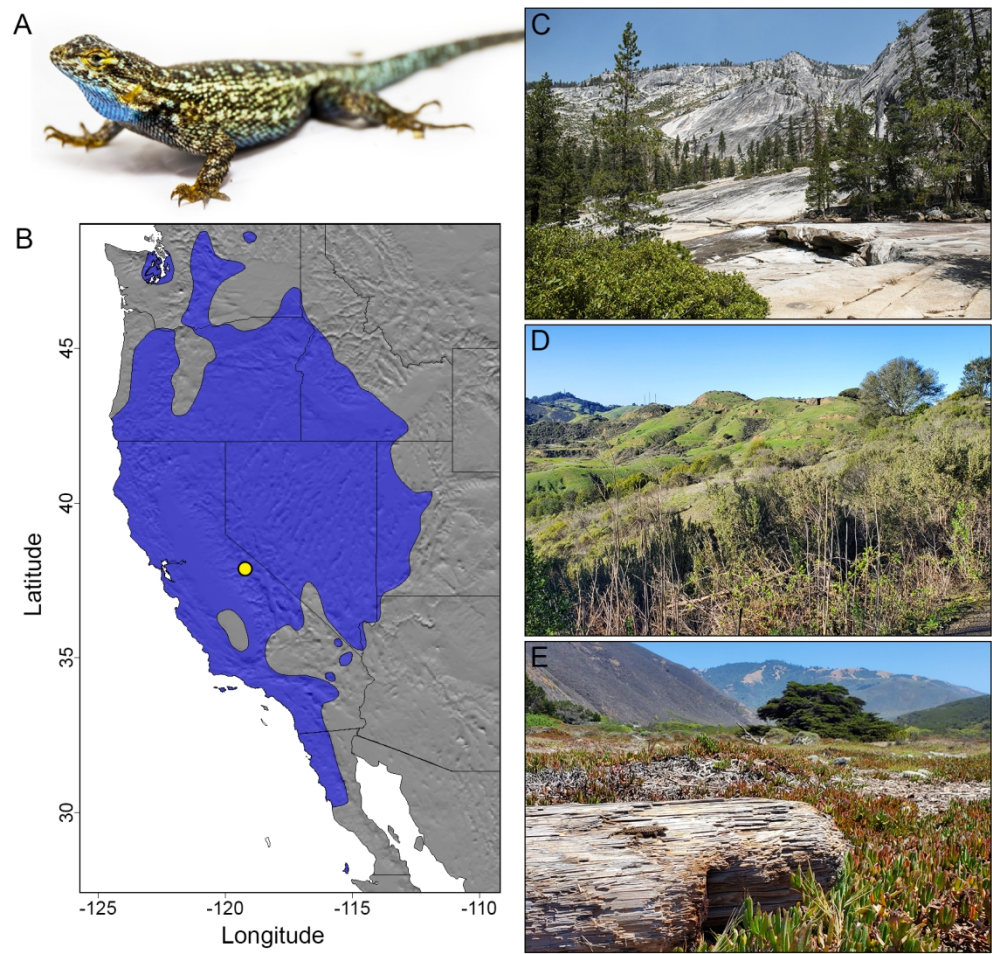


Figure 1. Photo of the western fence lizard (*Sceloporus occidentalis*) used to generate the reference genome (A) and a range map (B) for *S. occidentalis* based on data from the Global Assessment of Reptile Distributions (Roll et al. 2017) with the location for the genome specimen indicated by the yellow circle. Western fence lizards occupy habitat ranging from montane mixed conifer forest (C) to open woodland and grassland (D) to coastal scrub (E). Note the basking *S. occidentalis* on the log in the foreground of panel E.

381x365mm (118 x 118 DPI)

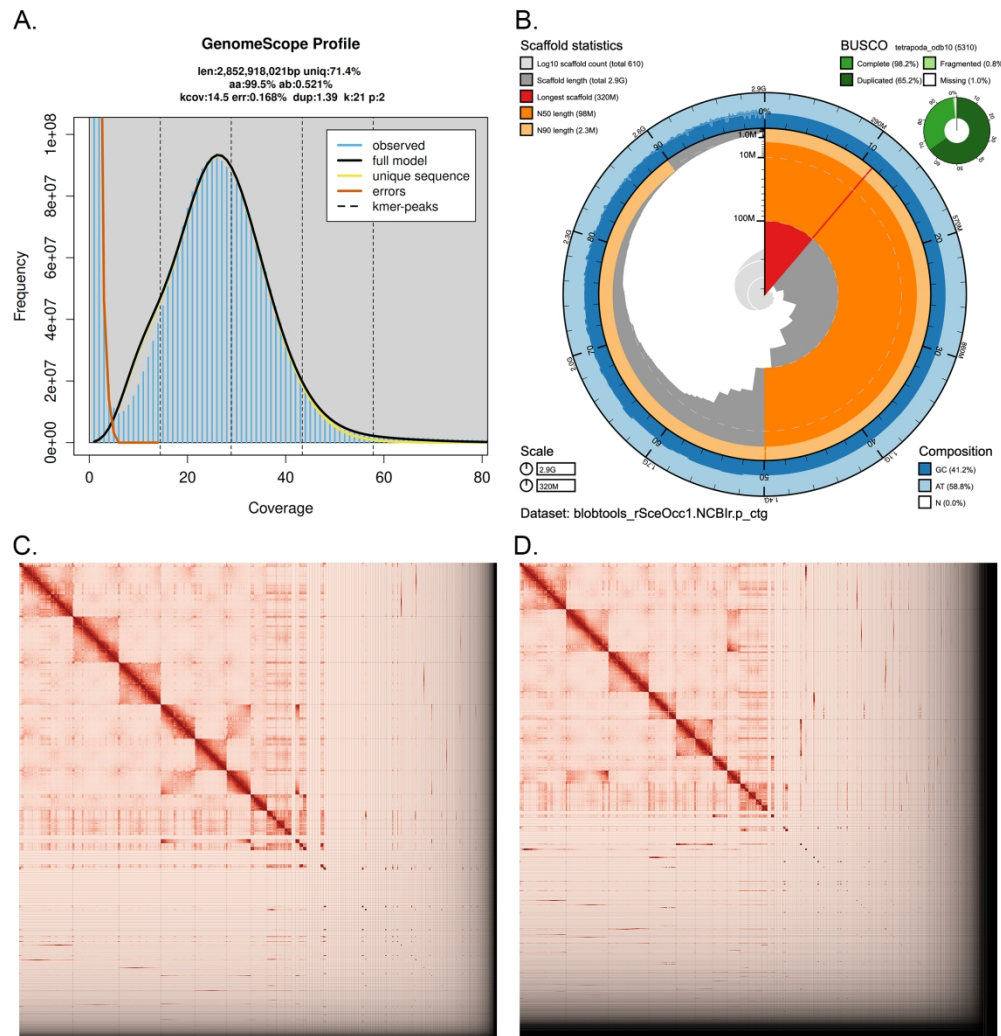


Figure 2. Visual overview of genome assembly metrics. A) K-mer spectrum output generated from PacBio HiFi data without adapters using GenomeScope2.0. B) BlobToolKit Snail plot showing a graphical representation of the quality metrics presented in Table 2 for the *S. occidentalis* primary assembly (rSceOcc1). 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-D) 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 3-D space to contiguous linear organization. Each cell in the contact map corresponds to sequencing data supporting the linkage (or join) between two such regions. Scaffolds are separated by black lines, and higher density corresponds to higher levels of fragmentation.

548x578mm (118 x 118 DPI)