



Genome Resources

A Reference Genome Assembly of the Bobcat, *Lynx rufus*

Meixi Lin¹, Merly Escalona², Ruta Sahasrabudhe, Oanh Nguyen, Eric Beraut, Michael R. Buchalski^{3*}, and Robert K. Wayne^{4*}

From the Department of Ecology and Evolutionary Biology, University of California, Los Angeles, CA 90095, USA (Lin and Wayne); Department of Biomolecular Engineering, University of California, Santa Cruz, Santa Cruz, CA 95064, USA (Escalona); DNA Technologies and Expression Analysis Core Laboratory, Genome Center, University of California, Davis, CA 95616, USA (Sahasrabudhe and Nguyen); Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, Santa Cruz, CA 95064, USA (Beraut); and Wildlife Genetics Research Unit, Wildlife Health Laboratory, California Department of Fish and Wildlife, Sacramento, CA 95834, USA (Buchalski).

*These authors contributed equally to this study.

Address correspondence to M. Lin at the address above, or e-mail: meixilin@ucla.edu

Corresponding Editor: Heath Blackmon

Abstract

The bobcat (*Lynx rufus*) is a medium-sized carnivore well adapted to various environments and an indicator species for landscape connectivity. It is one of the 4 species within the extant *Lynx* genus in the family *Felidae*. Because of its broad geographic distribution and central role in food webs, the bobcat is important for conservation. Here we present a high-quality de novo genome assembly of a male bobcat located in Mendocino County, CA, as part of the California Conservation Genomics Project (CCGP). The assembly was generated using the standard CCGP pipeline from a combination of Omni-C and HiFi technologies. The primary assembly comprises 76 scaffolds spanning 2.4 Gb, represented by a scaffold N50 of 142 Mb, a contig N50 of 66.2 Mb, and a BUSCO completeness score of 95.90%. The bobcat genome will be an important resource for the effective management and conservation of this species and comparative genomics exploration.

Key words: California Conservation Genomics Project, CCGP, carnivores, comparative genomics, felidae, long-read assembly

The bobcat (*Lynx rufus*) is one of the most adaptable and widespread carnivores in the Western Hemisphere (Figure 1A, Riley et al. 2003; Reding et al. 2012). They prefer rocky terrain in brushy forest or chaparral, but are also habitat generalists that can persist in anthropogenically altered areas (Figure 1B, Ahlborn and White 1990). For these reasons, the bobcat is an exemplary study species for functional landscape connectivity, urbanization effects, and local adaptations (Smith et al. 2020). With a large home range and central role in the food web, they are also considered an umbrella species for conserving diverse ecological communities (Kozakiewicz et al. 2019).

The bobcat shares a common ancestor with 3 other species in the *Lynx* genus, the Canada lynx (*Lynx canadensis*), the Eurasian lynx (*L. lynx*), and the Iberian lynx (*L. pardinus*), that diverged approximately 3.2 million years ago (Figure 1C, Johnson et al. 2006). Currently, the genomes for the Canada lynx (GCF_007474595.2; scaffold N50 = 147 Mb) and the Iberian lynx (GCA_900661375.1; scaffold N50 = 1.5 Mb) are available on NCBI Genbank (Supplementary Table 1). These 4 lynx species vary greatly in their ecological traits and demographic histories, as well as abundance and conservation status (Broderick 2020). Obtaining a high-quality bobcat reference genome will enable comparative genomics analyses in this lineage.

In California, the bobcat is a native mesocarnivore species that is crucial for ecosystem health (Ahlborn and White 1990). Regional studies using microsatellite or RADseq (Restriction site Associated DNA Sequencing) markers showed that habitat fragmentation, disease transmission, and rodenticide exposure increasingly pose threats to urban bobcats in southern California (Serieys et al. 2015; Fraser et al. 2018; Kozakiewicz et al. 2019). However, studies at a broader geographic scale examining the entire genome are still lacking. A bobcat reference genome would expand the genetic toolkit available to wildlife scientists responsible for the conservation and management of bobcat, both in California and throughout its native range.

The California Conservation Genomics Project (CCGP) is generating a genomics variation database for hundreds of species with broad statewide distributions to help guide conservation of species and ecosystems under anthropogenic changes (Shaffer et al. 2022). As one of the study species focused on by the CCGP, here we present a high-quality de novo genome assembly of the bobcat, with high contiguity, base accuracy, and minimal gaps. The assembly derives from genomic DNA extracted from fresh whole blood samples taken from a male bobcat that was treated at a wildlife rehabilitation facility. With high-molecular-weight (HMW) DNA, we leveraged the advantages of

Received March 14, 2022; Accepted June 10, 2022

© The American Genetic Association. 2022.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

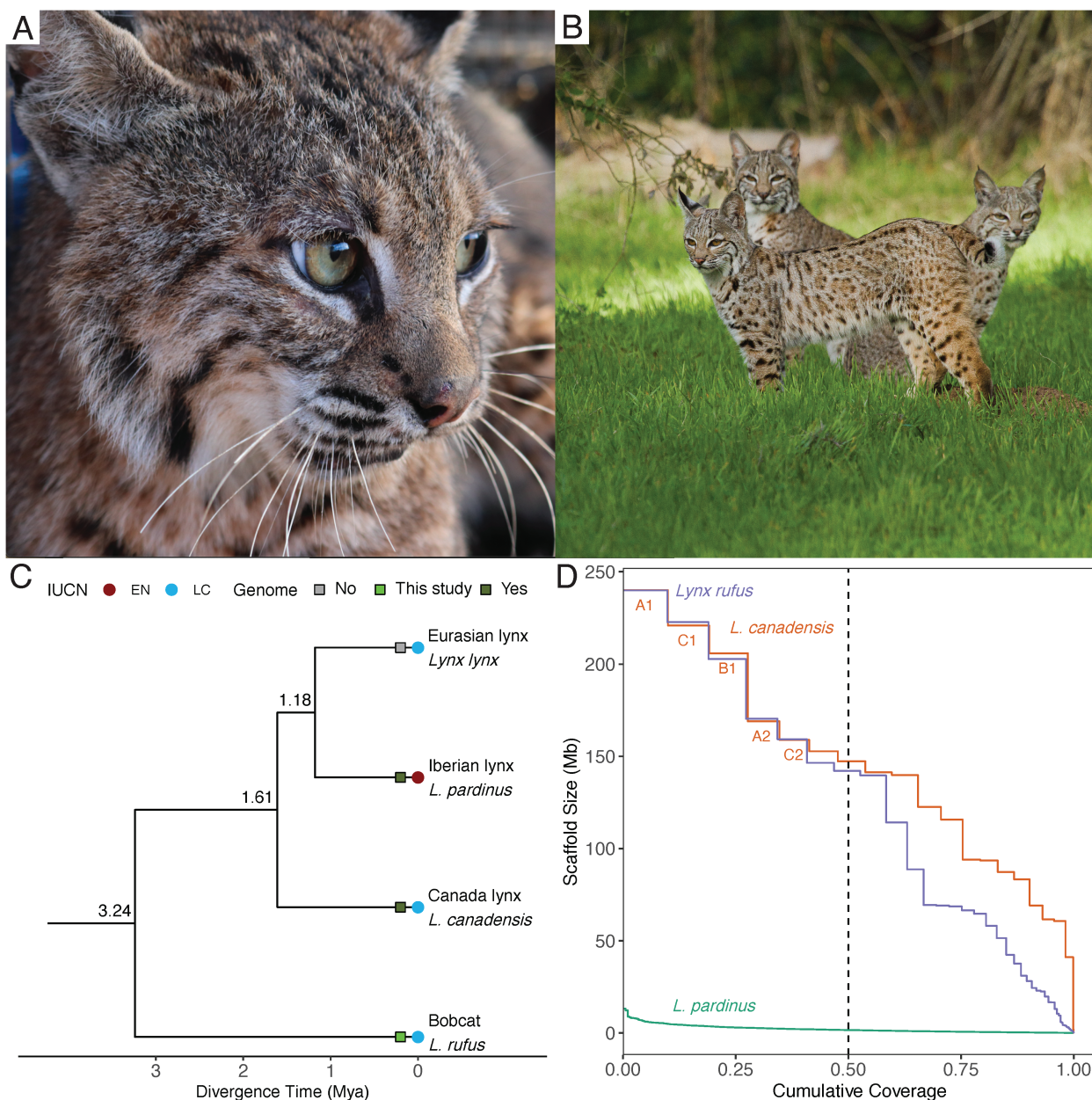


Figure 1. A bobcat reference genome assembly. **(A)** A bobcat, *Lynx rufus* (photograph credit: Laurel Serieys). **(B)** Representative habitats for bobcats (photograph credit: Barry Rowan). **(C)** Phylogenetic relationships in the *Lynx* genus. The IUCN Red List status (EN, Endangered; LC, Least Concern) and genome assembly availability (yes: available on NCBI, no: unavailable on NCBI) are denoted. Divergence time estimates are in units of million years ago (Mya, Johnson et al. 2006). **(D)** NGx plot comparing the 3 available *Lynx* genome assemblies. This plot shows the x fraction of genome assembly that is represented by scaffolds of at least y Mb. The N50 value is represented by the dashed vertical line. Our bobcat assembly (purple) has similar scaffold-level contiguity with the Canada lynx assembly (orange). The Iberian lynx assembly (green) has lower scaffold-level contiguity. The names for 5 longest Canada lynx chromosomes are annotated (See online version for color figure).

Omni-C proximity-ligation technologies and PacBio long-read sequencing to generate a high-quality assembly with chromosome-length scaffolds that is comparable to, or better than, existing lynx reference genomes (Abascal et al. 2016; Rhie et al. 2021). The bobcat assembly we present here has a total length of 2.44 Gb, a scaffold N50 of 142 Mb, and a contig N50 of 66.2 Mb. The bobcat genome will provide reference for high-resolution mapping of short-read data and genomic variation discovery in ongoing CCGP landscape genomics surveys and serve as a useful resource for comparative analyses.

Methods

Biological Materials

Whole blood was sampled from a male bobcat admitted to a wildlife rehabilitation facility for treatment of injuries sustained during a vehicle collision. The sample was collected under a Memorandum of Understanding with the California Department of Fish and Wildlife per CCR Title 14, Section 679. This male bobcat was found near Redwood Valley, Mendocino County (GPS: 39.26564 N, 123.15892 W, WGS84), CA. Whole blood samples were drawn in EDTA

blood collection tubes, refrigerated overnight, flash-frozen in liquid nitrogen and transferred on dry ice to the sequencing facilities within 24 hr of collection. Samples were then stored at -80°C until DNA extraction and sequencing. A voucher subsample is stored in the CCGP archive at the University of California–Los Angeles at -80°C .

Nucleic Acid Library Preparation and DNA Sequencing

Pacific Biosciences HiFi Library Preparation and Sequencing

HMW genomic DNA (gDNA) was isolated from whole blood preserved in EDTA. Three milliliters of RBC lysis solution (Qiagen Cat No. 158445) was added to 1 ml of whole blood and the reaction was incubated at room temperature for 5 min. The sample was centrifuged at $2000 \times g$ for 2 min to pellet white blood cells. The supernatant was discarded and 2 ml of lysis buffer containing 10 mM Tris–HCl pH 8.0, 25 mM EDTA, 0.5% (w/v) sodium dodecyl sulfate, and 100 $\mu\text{g}/\text{ml}$ Proteinase K was added to the cell pellet. The reaction was incubated at room temperature until the solution was homogenous. The lysate was then treated with 20 $\mu\text{g}/\text{ml}$ RNase A at 37°C for 30 min and cleaned with equal volumes of phenol/chloroform using phase lock gels (Quantabio Cat No. 2302830). The DNA was precipitated by adding 0.4 \times volume of 5 M ammonium acetate and 3 \times volume of ice cold ethanol. The DNA pellet was washed twice with 70% ethanol and resuspended in an elution buffer (10 mM Tris, pH 8.0). Purity of gDNA was measured on a NanoDrop 1000 spectrophotometer (Thermo Scientific, Waltham, MA) using the 260/280 and 260/230 ratios. The gDNA sample with a 260/280 ratio between 1.8 and 2.0 and a 260/230 ratio no less than 2.0 was considered pure (Pacific Biosciences 2021). The integrity of the HMW gDNA was verified on a Femto pulse system (Agilent Technologies, Santa Clara, CA).

The HiFi SMRTbell library was constructed using the SMRTbell Express Template Prep Kit v2.0 (Pacific Biosciences–PacBio; Menlo Park, CA; Cat. No. 100-938-900) according to the manufacturer's instructions. This entailed HMW gDNA shearing to a target DNA size distribution between 15 and 20 kb. The sheared gDNA was concentrated using 0.45 \times of AMPure PB beads (PacBio, Cat. No. 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 \times Ampure PB beads (PacBio, Cat. No. 100-265-900) for size selection using the BluePippin system (Sage Science, Beverly, MA; Cat No. BLF7510) to collect fragments greater than 9 kb. The 15–20 kb average HiFi SMRTbell library was sequenced at the University of California–Davis DNA Technologies Core (Davis, CA) using three 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 according to the manufacturer's protocol with slight modifications. Briefly, chromatin was fixed in

place in the nucleus. Fixed chromatin was digested with DNase I then extracted. 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. A sequencing library was generated using the NEB Ultra II DNA Library Prep kit (New England Biolabs, Ipswich, MA) with an Illumina compatible γ -adaptor. Biotin-containing fragments were then captured using streptavidin beads. The post-capture 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 at Vincent J. Coates Genomics Sequencing Lab (Berkeley, CA) on an Illumina NovaSeq platform (Illumina, San Diego, CA) to generate approximately 100 million paired end 150-bp reads per GB of genome size.

Genome Assembly

Nuclear Genome Assembly

We assembled the genome of the bobcat following the CCGP assembly protocol Version 3.0, an improvement from Todd et al. (2022). The main difference between versions is the use of an updated version of the *de novo* assembler HiFiasm [Version 0.16.1-r375] (Cheng et al. 2021, see Table 1 for assembly pipeline and relevant software). The final output corresponds to a diploid assembly that consists of 2 pseudo-haplotypes (primary and alternate). The primary assembly is more complete and consists of longer phased blocks. The alternate consists of haplotigs (contigs of clones with the same haplotype) in heterozygous regions and is not as complete and more fragmented. Given the characteristics of the latter, the alternate assembly cannot be considered on its own but as a complement of the primary assembly (<https://lh3.github.io/2021/04/17/concepts-in-phased-assemblies>; <https://www.ncbi.nlm.nih.gov/grc/help/definitions/>)

We removed remnant adapter sequence from the PacBio HiFi dataset using HiFiAdapterFilt [Version 1.0] (Sim 2021) and generated the initial diploid assembly with the filtered PacBio reads using HiFiasm. Next, we identified sequences corresponding to haplotypic duplications and contig overlaps on the primary assembly with *purge_dups* [Version 1.2.6] (Guan et al. 2020) and transferred them to the alternate assembly. We scaffolded both assemblies using the Omni-C data with SALSA [Version 2.2] (Ghurye et al. 2017, 2019).

The primary assembly was manually curated by generating and analyzing Omni-C contact maps and breaking the assembly where major misassemblies were found. No further joins were made after this step. To generate the contact maps, we aligned the Omni-C data against the corresponding reference with *bwa mem* [Version 0.7.17-r1188, options-SSP] (Li 2013), identified ligation junctions, and generated Omni-C pairs using *pairtools* [Version 0.3.0] (Goloborodko et al. 2019). We generated a multi-resolution Omni-C matrix with *cooler* [Version 0.8.10] (Abdennur and Mirny 2020) and balanced it with *hicExplorer* [Version 3.6] (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.

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

Assembly	Software	Version
Filtering PacBio HiFi adapters	HiFiAdapterFilt https://github.com/sheinasim/HiFiAdapterFilt	Commit 64d1c7b
K-mer counting	Meryl	1
Estimation of genome size and heterozygosity	GenomeScope	2
De novo assembly (contiging)	HiFiasm	0.16.1-r375
Long-read, genome–genome alignment	minimap2	2.16
Remove low-coverage, duplicated contigs	purge_dups	1.2.6
Scaffolding		
Omni-C mapping for SALSA	Arima Genomics mapping pipeline https://github.com/ArimaGenomics/mapping_pipeline	Commit 2e74ea4
Omni-C Scaffolding	SALSA	2
Gap closing	YAGCloser https://github.com/merlyescalona/yagcloser	Commit 20e2769
Omni-C Contact map generation		
Short-read alignment	bwa	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	3.6
Contact map visualization	HiGlass PretextMap PretextView PretextSnapshot	2.1.11 0.1.4 0.1.5 0.0.3
Organelle assembly		
Mitogenome assembly	MitoHiFi	2 Commit c06ed3e
Genome quality assessment		
Basic assembly metrics	QUAST	5.0.2
Assembly completeness	BUSCO Mercury	5.0.0 1
Contamination screening		
Local alignment tool	BLAST+	2.10
General contamination screening	BlobToolKit	2.3.3

We closed the remaining gaps generated during scaffolding with the PacBio HiFi reads and YAGCloser [commit 20e2769] (<https://github.com/merlyescalona/yagcloser>). We then checked for contamination using the BlobToolKit Framework [Version 2.3.3] (Challis et al. 2020). Finally, we trimmed remnants of sequence adaptors and mitochondrial contamination based on NCBI contamination screening.

Mitochondrial Genome Assembly

We assembled the mitochondrial genome of the bobcat from the PacBio HiFi reads using the reference-guided pipeline MitoHiFi (<https://github.com/marcelauliano/MitoHiFi>) (Allio et al. 2020). The mitochondrial sequence of *Lynx lynx* (MH706704.1) 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+ [Version 2.10] (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 ($k = 21$) from the PacBio HiFi reads using meryl [Version 1] (<https://github.com/marbl/meryl>). The generated k-mer database was then used in GenomeScope2.0 [Version 2.0] (Ranallo-Benavidez et al. 2020) to estimate genome features including sequencing error, genome size, heterozygosity, and repeat content. To obtain general contiguity metrics, we ran QUAST [Version 5.0.2] (Gurevich et al. 2013). To evaluate genome quality and completeness we used BUSCO [Version 5.0.0] (Simão et al. 2015; Seppey et al. 2019) with the mammalia database (mammalia_odb10) that contains 9226 genes. Assessment of base-level accuracy (QV) and k-mer completeness was performed using the previously generated meryl database and mercury (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).

Assembly Comparisons

We compared basic statistics with the other 2 existing nuclear assemblies and 3 mitochondrial assemblies in the *Lynx* genus

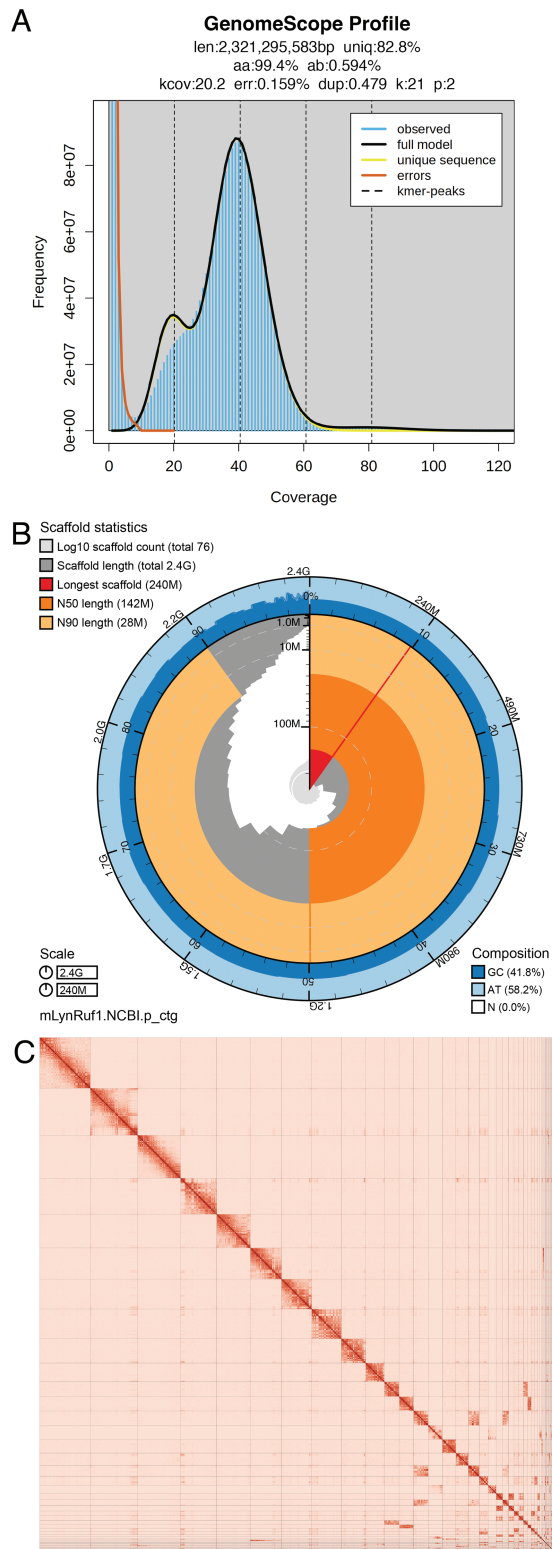


Figure 2. Visual overview of genome assembly metrics. **(A)** K-mer spectra output generated from PacBio HiFi data without adapters using GenomeScope2.0. The bimodal pattern observed corresponds to a diploid genome. K-mers covered at lower coverage and lower frequency correspond to differences between 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 *Lynx rufus* primary assembly (mLynRuf1). The plot circle represents the full size of the assembly. From the inside-out, the central plot covers

(Figure 1C). For nuclear assemblies, we downloaded the *Lynx canadensis* genome RefSeq assembly (GCF_007474595.2_mLynCan4.pri.v2; accessed 13 Aug 2021) and the *Lynx pardinus* assembly (GCA_900661375.1_LYP1.0; accessed 16 February 2022). To compare basic statistics in the nuclear assemblies, we compiled information from the NCBI Genome Assembly Reports and individual publications (Supplementary Table 1, Abascal et al. 2016; Rhie et al. 2021). To standardize the BUSCO scores, we repeated the BUSCO analyses described above for the other assemblies (Supplementary Table 2). The divergence time plot was generated using the ggtree package [Version 2.0.4] (Yu et al. 2017) in R [Version 3.6.2] (R Core Team 2019). The coverage by scaffold length (NGx) plot was generated based on the scaffold lengths in the NCBI Genome Assembly Reports for each species using ggplot2 [Version 3.3.2] (Wickham 2016) in R.

For mitochondrial assemblies, we downloaded the sequences for GenBank accessions: CM017348.2 (*Lynx canadensis*), MH706704.1 (*Lynx lynx*), and NC_028319.1 (*Lynx pardinus*) on 20 February 2022. The base compositions and sequence lengths were summarized using biopython [Version 1.79] (Cock et al. 2009).

To count available whole genome assemblies in the Felidae, we queried the NCBI Assembly database using the Felidae taxonomy ID on 21 February 2022 (<https://www.ncbi.nlm.nih.gov/assembly/?term=txid9681%5BOrganism%3Aexp%5D>). The assembly species names were matched to the Felidae taxonomy described in Kitchener et al. (2017).

Results

Nuclear Assembly

We generated a de novo nuclear genome assembly of the bobcat (mLynRuf1) using 247.6 million read pairs of Omni-C data and 6.7 million PacBio HiFi reads. The latter yielded ~40 fold coverage (N50 read length 14 593 bp; minimum read length 45 bp; mean read length 14 504 bp; maximum read length of 52 209 bp) based on the final assembled genome size of 2.4 Gb (Figure 2A). Assembly statistics are reported in tabular and graphical form in Table 2 and Figure 2B, respectively.

The primary assembly consists of 76 scaffolds spanning 2.4 Gb with contig N50 of 66.2 Mb, scaffold N50 of 142.1 Mb, largest contig of 202.7 Mb, and largest scaffold of 239.8 Mb. Using BlobToolKit and BLAST+, we identified and removed 1 contig from the primary assembly corresponding to mitochondrial contamination, and 7 contigs from the alternate

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). **(C)** The Omni-C contact map for the primary genome assembly generated with PretextSnapshot. Omni-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 two of such regions. Scaffolds are separated by black lines, and higher density corresponds to higher levels of fragmentation (See online version for color figure).

Table 2. Sequencing and assembly statistics, and accession numbers

BioProjects and vouchers	CCGP NCBI BioProject		PRJNA720569			
	Genera NCBI BioProject		PRJNA765621			
	Species NCBI BioProject		PRJNA777191			
	NCBI BioSample		SAMN23391104			
	Specimen identification		CCGP_SWC_20201006			
Genome sequence	NCBI Genome accessions		Primary		Alternate	
	Assembly accession		GCA_022079265.1		GCA_022079275.1	
	Genome sequences		JAJSDN000000000		JAJSDO000000000	
Sequencing data	PacBio HiFi reads	Run	3 PACBIO_SMRT (Sequel II) runs: 6.7 M spots, 97.2 G bases, 65.5Gb			
		Accession	SRR17978068			
	Omni-C Illumina reads	Run	2 Illumina NovaSeq 6000 runs: 247.6 M spots, 74.8 G bases, 24.9 Gb			
		Accession	SRR17978066-67			
Genome assembly quality metrics	Assembly identifier (quality code ^a)		mLynRuf1 (7.8.Q66)			
	HiFi Read coverage ^b		40×			
			Primary		Alternate	
	Number of contigs		100		72 828	
	Contig N50 (bp)		66 217 191		97 441	
	Longest contigs		202 776 522		1 850 898	
	Number of scaffolds		76		68 112	
	Scaffold N50 (bp)		142 134 035		107 693	
	Largest scaffold		239 891 529		11 854 743	
	Size of final assembly (bp)		2 439 256 471		3 320 187 595	
	Gaps per Gbp		8		1420	
	Indel QV (frame shift)		47.15		47.15	
	Base pair QV		66.5		58.25	
			Full assembly = 60.19			
	k-mer completeness		96.31		88.12	
			Full assembly = 99.74			
	BUSCO		C	S	D	F
Completeness (mammalia)		P ^c 95.90%	95.20%	0.70%	1.20%	2.90%
<i>N</i> = 9226		A ^c 81.10%	74.60%	6.50%	5.10%	13.80%
Organelles		1 Complete mitochondrial sequence			CM039064.1	

^aAssembly quality code *x.y.Q* derived notation, from [Rhie et al. \(2021\)](#). *x* = log10[contig NG50]; *y* = log10[scaffold NG50]; *Q* = Phred base accuracy QV (Quality value). 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; bp, base pairs.

^bRead coverage has been calculated based on a genome size of 2.4 Gb.

^cP(primary) and (A)lternate assembly values.

assembly, 6 contigs corresponding to mitochondrial contamination, and 1 contig to an arthropod contaminant. The Omni-C contact map suggests that the primary assembly is highly contiguous ([Figure 2C](#)). As expected, the alternate assembly, which consists of sequence from heterozygous regions, is less contiguous ([Supplementary Figure 1](#)). Because the primary assembly is not fully phased, we have deposited scaffolds corresponding to the alternate haplotype in addition to the primary assembly.

The final genome size (2.4 Gb) is close to the estimated values from the Genomescope2.0 k-mer spectra. The k-mer spectrum output shows a bimodal distribution with 2 major peaks, at

~19- and ~38-fold coverage, where peaks correspond to homozygous and heterozygous states, respectively ([Figure 2A](#)).

Based on PacBio HiFi reads, we estimated a 0.15% sequencing error rate and 0.59% nucleotide heterozygosity rate. The assembly has a BUSCO completeness score of 95.9% using the mammalia gene set, a per base quality (QV) of 66, a k-mer completeness of 96.3, and a frameshift indel QV of 47.15.

Mitochondrial Assembly

The mitochondrial genome assembled with MitoHiFi has a final size of 17 097 bp. The base composition of the final assembly version is A = 32.43%, C = 26.64%, G = 14.24%,

T = 26.69%, and consists of 22 transfer RNAs and 13 protein coding genes. Within the *Lynx* genus, the mitochondrial genome size is conserved (16 806–17 097 bp). The mitochondrial base compositions vary little across species as well (A = 32.31–32.43%, C = 26.64–27.06%, G = 14.16–14.29%, T = 26.35–26.69%; [Supplementary Table 3](#)).

Discussion

Here we presented a high-quality bobcat reference genome assembly, with some scaffold lengths reaching chromosome levels. The 5 longest scaffolds in the bobcat assembly have almost identical lengths compared with the assigned A1, C1, B1, A2, and C2 chromosomes in the Canada lynx assembly ([Figure 1D](#)). This bobcat assembly is highly continuous, complete, and accurate. With a contig N50 of 66 Mb, scaffold N50 of 142 Mb, total gap length of 2.4 kb, and a base pair QV of 66, our assembly greatly exceeds the best available standards of a minimum contig N50 of 1Mb, scaffold N50 of 10 Mb, and QV of 40 proposed by the Vertebrate Genome Project (VGP; [Rhie et al. 2021](#)).

Compared with the other 2 available *Lynx* assemblies for Canada lynx (LYCA) and Iberian lynx, our bobcat assembly (LYRU) is of similar quality to the VGP Canada lynx assembly, which also utilized both long- and short-read sequences ([Figure 1D](#); [Supplementary Table 1](#)). We achieved a higher accuracy (QV = 66) compared with the Canada lynx assembly (QV = 36.8), a slightly higher BUSCO completeness score (95.9% for LYRU and 94.4% for LYCA; [Supplementary Table 2](#)), less total gap length (2.4 kb for LYRU and 2.8 Mb for LYCA) and equivalent k-mer completeness (96.3% for LYRU and 96.4% for LYCA). The Canada lynx assembly generated chromosome assignments, which are not included in this current bobcat assembly release. Both bobcat and Canada lynx genomes were annotated using the NCBI Eukaryotic Genome Annotation Pipeline ([Supplementary Table 1](#)). BUSCO analysis of gene annotations for our bobcat assembly using the carnivora_odb10 lineage dataset showed a 98.5% completeness, suggesting a high annotation quality (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Lynx_rufus/100/). Both bobcat and Canada lynx genomes are superior in assembly metrics compared with the Iberian lynx assembly that was generated using only short-read sequencing techniques ([Figure 1D](#)).

This bobcat assembly will provide resources for comparative genomic studies within the *Lynx* lineage, and more broadly the *Felidae* family. Of the 41 living felid species in 8 *Felidae* lineages ([Kitchener et al. 2017](#)), 17 species have at least 1 genome assembly available in NCBI ([Supplementary Table 4](#)). Assembled genome sizes in the *Lynx* lineage, measured in total genome assembly length, are approximately 2.4 Gb for all 3 existing assemblies ([Supplementary Table 1](#)). At a larger scale, the genome sizes in the *Felidae* family are also conserved (2.30 Gb for *Panthera leo*, GCF_018350215.1 to 2.58 Gb for *Panthera pardus*, GCF_001857705.1; [Supplementary Table 4](#)). The assembled *L. rufus* genome size is smaller than the flow cytometry measured size of 2.92 Gb for the *Lynx lynx* ([Vinogradov 1998](#); [Gregory 2005](#)), a pattern observed in other species possibly caused by the repetitive regions ([Elliott and Gregory 2015](#)). Species within the *Lynx* lineage vary in abundance and conservation status, which is reflected in the nucleotide heterozygosity. The more abundant bobcat and

Canada lynx have 0.59 and 0.19% heterozygosity in their assemblies ([Rhie et al. 2021](#)), while only 0.01% heterozygosity was reported for the endangered Iberian lynx ([Abascal et al. 2016](#)).

In addition to evolutionary studies, the bobcat reference genome will be an essential resource for genetics-informed conservation management. Currently, a bobcat-hunting ban is in place in California until 2025, at which time the Fish and Game Commission must re-evaluate the appropriateness of a hunting season based on the best available science ([California Assembly Bill No. 1254, 2019](#)). To support this evaluation, the California Statewide Bobcat Population Monitoring project is underway to assess population status ([CDFW 2021](#)). The bobcat genome will provide reference for ongoing CCGP whole genome resequencing projects that aim to identify statewide Management Units, evaluate genomic health, and assess the outcomes of various hunting scenarios through genomics-informed simulations. Across its geographic range from southern Canada to Mexico ([Kelly et al. 2016](#)), researchers have studied bobcats to characterize patterns of genetic variation on a continental scale ([Reding et al. 2012](#); [Broderick 2020](#)) and to assess impacts of habitat fragmentation on gene flow ([Serieys et al. 2015](#); [Janecka et al. 2016](#)), as well as urbanization associated disease and toxins ([Fraser et al. 2018](#); [Kozakiewicz et al. 2020](#)). The availability of a high-quality genome assembly will further advance research topics such as these as well.

In summary, this highly contiguous, complete, and accurate assembly for bobcat is a part of the larger goal of the California Conservation Genomics Project to build the most comprehensive conservation genomics dataset known to date. The availability of such high-quality assemblies will serve as an important tool for both fundamental evolutionary studies and conservation applications.

Supplementary Material

Supplementary data are available at *Journal of Heredity* online.

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). Sample collection was supported by funding from the U.S. Fish and Wildlife Service, Wildlife Restoration Act (grant no. P1580009).

Acknowledgments

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. We thank Dr. Devaughn Fraser, Dr. Laurel Serieys, Dr. Kirk Lohmueller, Dr. Brad Shaffer, and Dr. Erin Toffelmier for inputs on study design; Mr. Daniel R. Oliveira, Dr. Courtney Miller, and Ms. Tara Luckau for coordination on sample submissions; and Mr. Barry Rowan and Dr. Laurel Serieys for sharing bobcat photos in Figure 1. We thank the staff at Sonoma Wildlife Care and California's wildlife rehabilitation facilities for their generous help with providing the samples. This work used computational and storage services associated with the Hoffman2 Shared Cluster provided by UCLA Institute for Digital Research and Education's Research Technology Group.

Data Availability

Data generated for this study are available under NCBI BioProject PRJNA777191. Raw sequencing data for sample CCGP_SWC_20201006 (NCBI BioSample SAMN23391104) are deposited in the NCBI Short Read Archive (SRA) under SRR17978068 for PacBio HiFi sequencing data and SRR17978066-67 for Omni-C Illumina Short read sequencing data. GenBank accessions for both primary and alternate assemblies are GCA_022079265.1 and GCA_022079275.1; and for genome sequences JAJSDN000000000 and JAJSDO000000000. The NCBI RefSeq accession corresponding to the primary assembly GCA_022079265.1 is GCF_022079265.1. The GenBank organelle genome assembly for the mitochondrial genome is CM039064.1. Assembly scripts and other data for the analyses presented can be found at the following GitHub repository: https://www.github.com/ccgproject/ccgp_assembly. The scripts for genome assembly comparisons can be found at the following GitHub repository: https://github.com/meixilin/ccgp_bobcat_joh.

References

- Abascal F, Corvelo A, Cruz F, Villanueva-Cañas JL, Vlasova A, Marcet-Houben M, Martínez-Cruz B, Cheng JY, Prieto P, Quesada V, et al. 2016. Extreme genomic erosion after recurrent demographic bottlenecks in the highly endangered Iberian lynx. *Genome Biol.* 17:251.
- Abdennur N, Mirny LA. 2020. Cooler: scalable storage for hi-c data and other genomically labeled arrays. edited by Jonathan Wren. *Bioinformatics.* 36(1):311–316.
- Ahlborn G, White M. 1990. California's wildlife, bobcat. Available from: <https://nrm.dfg.ca.gov/FileHandler.ashx?DocumentID=2609&inline=1>.
- Allio R, Schomaker-Bastos A, Romiguier J, Prodocimi F, Nabholz B, Delsuc F. 2020. MitoFinder: efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. *Mol Ecol Resour.* 20(4):892–905.
- Broderick J. 2020. A genomic analysis of bobcat populations in North America with a comparison to the Canada Lynx: an assessment of local adaptation to unique ecoregions and phylogeography. Available from: <https://dsc.duq.edu/etd/1886>.
- California Assembly Bill No. 1254, 2019. Bill Text—AB-1254 bobcats: take prohibition: hunting season: management plan. 2019. Available from: https://leginfo.ca.gov/faces/billTextClient.xhtml?bill_id=2019020200AB1254.
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinf.* 10:421.
- CDFW. 2021. Science Institute News. CDFW Begins Statewide Bobcat Monitoring Project. 14 May 2021. <https://wildlife.ca.gov/Science-Institute/News/cdfw-begins-statewide-bobcat-monitoring-project>.
- Challis R, Richards E, Rajan J, Cochrane G, Blaxter M. 2020. BlobToolKit—interactive quality assessment of genome assemblies. *G3.* 10:1361–1374.
- Cheng H, Jarvis ED, Fedrigo O, Koepfli K-P, Urban L, Gemmill NJ, Li H. 2021. Robust haplotype-resolved assembly of diploid individuals without parental data. *ArXiv:2109.04785 [q-Bio]*, September, <http://arxiv.org/abs/2109.04785>.
- Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, et al. 2009. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics.* 25:1422–1423.
- Elliott TA, Ryan Gregory T. 2015. What's in a genome? The c-value enigma and the evolution of eukaryotic genome content. *Philos Trans R Soc B Biol Sci.* 370:20140331.
- Fraser D, Mouton A, Serieys LEK, Cole S, Carver S, Vandewoude S, Lappin M, Riley SPD, Wayne R. 2018. Genome-wide expression reveals multiple systemic effects associated with detection of anticoagulant poisons in bobcats (*Lynx rufus*). *Mol Ecol.* 27:1170–1187.
- Ghurye J, Pop M, Koren S, Bickhart D, Chin C-S. 2017. Scaffolding of long read assemblies using long range contact information. *BMC Genomics.* 18:527.
- Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, Phillippy AM, Koren S. 2019. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS Comput Biol.* 15:e1007273.
- Goloborodko A, Abdennur N, Venev S, Hbbrandao, Gfudenberg. 2019. Mirnylab/Pairtools v0.3.0. *Zenodo.* doi:10.5281/zenodo.2649383
- Gregory TR. 2005. Animal Genome Size Database. <http://www.genomesize.com>. Accessed 13 May 2022.
- Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. 2020. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics.* 36:2896–2898.
- Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics.* 29:1072–1075.
- Janecka JE, Tewes ME, Davis IA, Haines AM, Caso A, Blankenship TL, Honeycutt RL. 2016. Genetic differences in the response to landscape fragmentation by a habitat generalist, the bobcat, and a habitat specialist, the ocelot. *Conserv Genet* 17:1093–1108.
- Johnson WE, Eizirik E, Pecon-Slattery J, Murphy WJ, Antunes A, Teeling E, O'Brien SJ. 2006. The late Miocene radiation of modern Felidae: a genetic assessment. *Science.* 311(5757):73–77. doi:10.1126/science.1122277
- Kelly M, Morin D, Lopez-Gonzalez CA. 2016. *Lynx rufus*. The IUCN Red List of Threatened Species 2016: E.T12521A50655874. *International Union for Conservation of Nature.* doi:10.2305/IUCN.UK.2016-1.RLTS.T12521A50655874.en
- Kerpedjiev P, Abdennur N, Lekschas F, McCallum C, Dinkla K, Strobelt H, Lubner JM, et al. 2018. HiGlass: web-based visual exploration and analysis of genome interaction maps. *Genome Biol.* 19:125.
- Kitchener AC, Breitenmoser-Würsten C, Eizirik E, Gentry A, Werdelin L, Wilting A, Yamaguchi N, et al. 2017. A revised taxonomy of the Felidae: the final report of the Cat Classification Task Force of the IUCN Cat Specialist Group. Available from: <http://repository.si.edu/xmlui/handle/10088/32616>.
- Korlach J, Gedman G, Kingan SB, Chin C-S, Howard JT, Audet J-N, Cantin L, Jarvis ED. 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.
- Kozakiewicz CP, Burrige CP, Chris Funk W, Craft ME, Crooks KR, Fisher RN, Fountain-Jones NM, et al. 2020. Does the virus cross the road? Viral phylogeographic patterns among bobcat populations reflect a history of urban development. *Evol Appl.* 13:1806–1817.
- Kozakiewicz CP, Burrige CP, Chris Funk W, Salerno PE, Trumbo DR, Gagne RB, Boydston EE, Fisher RN, Lyren LM, Jennings MK, et al.

2019. Urbanization reduces genetic connectivity in bobcats (*Lynx rufus*) at both intra- and interpopulation spatial scales. *Mol Ecol.* 28:5068–5085.
- Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv:1303.3997* [q-Bio], May, <http://arxiv.org/abs/1303.3997>.
- Pacific Biosciences. 2021. Technical overview: ultra-low DNA input library preparation using SMRTbell Express Template Prep Kit 2.0. <https://www.pacb.com/wp-content/uploads/Ultra-Low-DNA-Input-Library-Preparation-Using-SMRTbell-Express-TPK-2.0-Customer-Training-01.pdf>. Accessed 13 May 2022.
- Ramírez F, Bhardwaj V, Arrigoni L, Lam KC, Grüning BA, Villavecis J, Habermann B, Akhtar A, Manke T. 2018. High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat Commun.* 9:189.
- Ranallo-Benavidez TR, Jaron KS, Schatz MC. 2020. GenomeScope 2.0 and smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun.* 11:1432.
- R Core Team. 2019. *R: a language and environment for statistical computing*. Vienna (Austria): R Foundation for Statistical Computing. Available from: <https://www.R-project.org/>.
- Reding DM, Bronikowski AM, Johnson WE, Clark WR. 2012. Pleistocene and ecological effects on continental-scale genetic differentiation in the bobcat (*Lynx rufus*). *Mol Ecol.* 21:3078–3093.
- Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W, Functamman A, Kim J, et al. 2021. Towards complete and error-free genome assemblies of all vertebrate species. *Nature.* 592:737–746.
- Rhie A, Walenz BP, Koren S, Phillippy AM. 2020. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* 21:245.
- Riley SPD, Sauvajot RM, Fuller TK, York EC, Kamradt DA, Bromley C, Wayne RK. 2003. Effects of urbanization and habitat fragmentation on bobcats and coyotes in Southern California. *Conserv Biol.* 17:566–576.
- Seppey M, Manni M, and Zdobnov EM. 2019. BUSCO: assessing genome assembly and annotation completeness. In: Kollmar M, editor. *Gene prediction. Methods in Molecular Biology*. New York (NY): Springer New York. p. 227–245.
- Serieys LEK, Lea A, Pollinger JP, Riley SPD, Wayne RK. 2015. Disease and freeways drive genetic change in urban bobcat populations. *Evol Appl.* 8:75–92.
- Shaffer HB, Toffelmier E, Corbett-Detig RB, Escalona M, Erickson B, Fiedler P, Gold M, Harrigan RJ, Hodges S, Luckau TK, et al. 2022. Landscape genomics to enable conservation actions: the California conservation genomics project. *J Hered.* 113: 577–588.
- Sim S. 2021. Sheinasim/HiFiAdapterFilter: first release (version v1.0.0). Zenodo. doi:10.5281/ZENODO.4716418
- Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics.* 31:3210–3212.
- Smith JG, Jennings MK, Boydston EE, Crooks KR, Ernest HB, Riley SPD, Serieys LEK, Sleater-Squires S, Lewison RL. 2020. Carnivore population structure across an urbanization gradient: a regional genetic analysis of bobcats in Southern California. *Landsc Ecol.* 35:659–674.
- Todd BD, Jenkinson TS, Escalona M, Beraut E, Nguyen O, Sahasrabudhe R, Scott PA, Toffelmier E, Wang IJ, Shaffer HB. 2022. Reference genome of the northwestern pond turtle, *Actinemys marmorata*. *J Hered.* doi:10.1093/jhered/esac021
- Vinogradov AE. 1998. Genome size and GC-percent in vertebrates as determined by flow cytometry: the triangular relationship. *Cytometry.* 31:100–109.
- Wickham H. 2016. *Ggplot2: elegant graphics for data analysis*. New York (NY): Springer-Verlag. Available from: <https://ggplot2.tidyverse.org>.
- Yu G, Smith D, Zhu H, Guan Y, Lam TTY. 2017. Ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol.* 8:28–36.