Title

A highly contiguous genome assembly for the California quail (Callipepla californica)

Permalink

https://escholarship.org/uc/item/38m80553

Journal

Journal of Heredity, 114(4)

ISSN

0022-1503

Authors

Benham, Phred M
Cicero, Carla
Escalona, Merly
et al.

Publication Date

2023-06-22

DOI

10.1093/jhered/esad008

Peer reviewed

Journal of Heredity, 2023, **114**, 418–427
https://doi.org/10.1093/jhered/esad008
Advance access publication 10 February 2023
Genome Resources

Genome Resources

A highly contiguous genome assembly for the California quail (*Callipepla californica*)

Phred M. Benham[1,2], Carla Cicero[1], Merly Escalona[3], Eric Beraut[4], Mohan P.A. Marimuthu[5], Oanh Nguyen[5], Michael W. Nachman[1,2], Rauri C.K. Bowie[1,2]

[1]Museum of Vertebrate Zoology, University of California Berkeley, Berkeley, CA, United States,
[2]Department of Integrative Biology, University of California Berkeley, Berkeley, CA, United States,
[3]Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, 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

Address correspondence to P.M. Benham at the address above, or e-mail: phbenham@gmail.com.

Corresponding Editor: Arun Sethuraman

Abstract

The California quail (*Callipepla californica*) is an iconic native bird of scrub and oak woodlands in California and the Baja Peninsula of Mexico. Here, we report a draft reference assembly for the species generated from PacBio HiFi long read and Omni-C chromatin-proximity sequencing data as part of the California Conservation Genomics Project (CCGP). Sequenced reads were assembled into 321 scaffolds totaling 1.08 Gb in length. Assembly metrics indicate a highly contiguous and complete assembly with a contig N50 of 5.5 Mb, scaffold N50 of 19.4 Mb, and BUSCO completeness score of 96.5%. Transposable elements (TEs) occupy 16.5% of the genome, more than previous Odontophoridae quail assemblies but in line with estimates of TE content for recent long-read assemblies of chicken and Peking duck. Together these metrics indicate that the present assembly is more complete than prior reference assemblies generated for Odontophoridae quail. This reference will serve as an essential resource for studies on local adaptation, phylogeography, and conservation genetics in this species of significant biological and recreational interest.

Key words: California Conservation Genomics Project, Odontophoridae, transposable elements, upland game bird

Introduction

The charismatic California quail (*Callipepla californica*) is the state bird of California. The "*chi-ca-go*" rally calls given by both sexes can be heard year-round from throughout its native distribution in the Baja peninsula of Mexico, California, southern Oregon, and western Nevada (Fig. 1; Calkins *et al.* 2014). Starting in the mid-1800s, successful introductions were carried out widely across western North America (e.g. Washington, Idaho, Montana, Utah), Chile, Argentina, New Zealand, Australia, Hawaii, and Corsica (Calkins *et al.* 2014). California quail are typically found in a variety of oak woodland, chaparral, and sage habitats within its native range. The species spans broad thermal and precipitation gradients, but in the most arid regions of southeastern California it is replaced by its sister species, the Gambel's quail (*Callipepla gambeli*). The 2 form a narrow hybrid zone in the Coachella valley in Riverside County, California (Zonana *et al.* 2019).

California and Gambel's quail comprise half the members of the genus *Callipepla*, a lineage within the family Odontophoridae. Most of the 33 Odontophoridae quail species occupy tropical forest environments, but *Callipepla* and its sister genus, *Colinus* bobwhites, inhabit more open and arid habitats (Hosner *et al.* 2015; Winkler *et al.* 2020; Salter et al. 2022). Within California quail, 5 subspecies are currently accepted (Calkins *et al.* 2014). *Callipepla c. californica* is the most widespread subspecies ranging from southern Oregon south through much of California to northern Baja California, Mexico. It is replaced south of central Baja California by *C. c. achrustera* and in northwestern California by *C. c. brunnescens*. *Callipepla c. canfieldae* is restricted to the Owen's valley of eastern California and *C. c. catalinensis* is native only to Catalina Island off the coast of southern California. Population genetic data are largely lacking for California quail, although an early allozyme analysis suggested that some subtle population structure may exist between populations in California and Baja California (Zink *et al.* 1987).

California quail are generally considered widespread and of least conservation concern. An exception is the endemic Catalina Island population (*C. c. catalinensis*), which is classified as a California Bird Species of Special Concern due to the narrowness of its distribution (Collins 2008). Quail populations across the state of California also have declined precipitously from historic highs in the 1800s. These

Received December 7, 2022; Accepted February 9, 2023

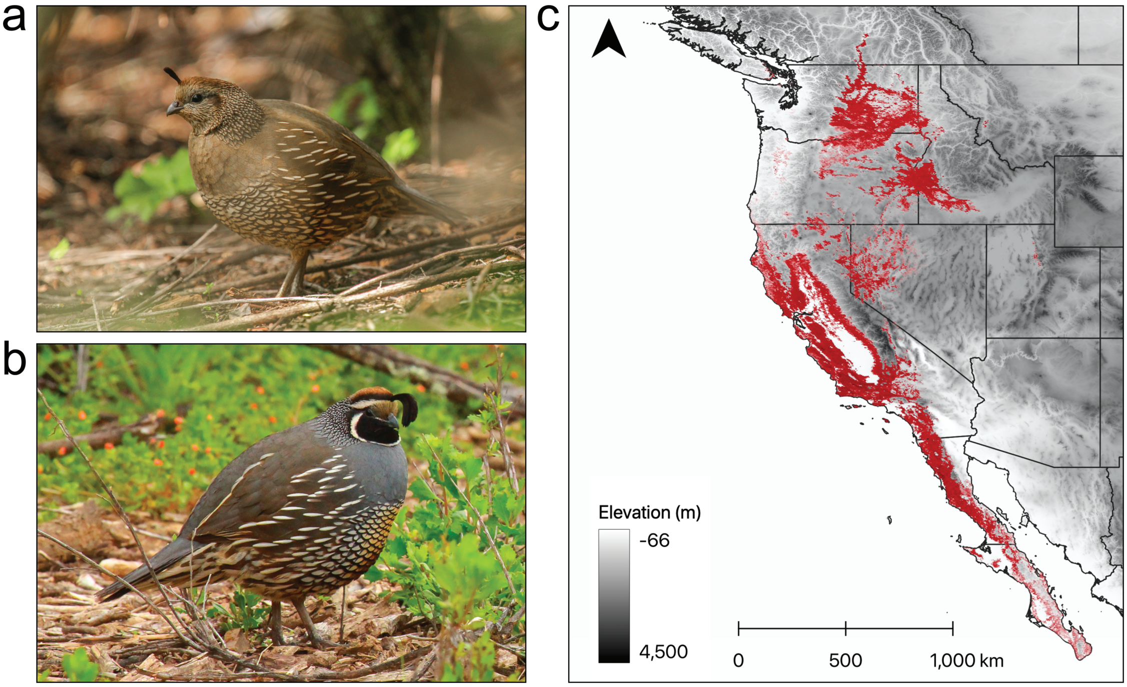

Fig. 1. Female (a) and male (b) California quail (*Callipepla californica*). (c) Distribution of California quail (red) across western North America. Map includes both native and introduced populations. Distribution data from eBird (https://science.ebird.org/). Photos taken at Blue Oak Ranch Reserve, Santa Clara County, California and courtesy of Jackie Childers.

declines were driven by unregulated hunting and trapping that supplied tens of thousands of quail annually to markets in San Francisco and Los Angeles (e.g. 177,366 quail in 1895; Grinnell *et al.* 1918; Leopold 1977). Bag limits and anti-trapping laws were established in 1901, but ongoing habitat degradation from overgrazing, mechanized agriculture, and urbanization have all contributed to a constricted quail distribution with lower densities than historically reported. Furthermore, persistent drought conditions and the increasing severity of wildfires in the western United States continue to pose significant threats to California quail across its distribution (Calkins *et al.* 2014). Finally, free-ranging (feral) cats also can reduce the likelihood of habitat occupancy by California quail, especially in urban environments (Iknayan *et al.* 2021).

In light of these historic and future threats to California quail populations, the species has been included in the California Conservation Genomics Project (CCGP; Shaffer *et al.* 2022). The goals of the CCGP are to better understand shared patterns of genetic diversity and population structure across over 150 California plant and animal species. A key component of each CCGP project is the generation of high-quality reference genomes for each species. Here, we report on the reference assembly of the California quail generated by the CCGP using PacBio HiFi long-read and Omni-C chromatin-proximity sequencing data. This represents the first long-read-based genome assembly for the family Odontophoridae. Along with high-quality reference genomes for the domestic chicken (*Gallus gallus*; Warren *et al.* 2017) and mallard duck (*Anas platyrhynchos*; Zhu *et al.* 2021) the California quail genome generated for the CCGP will serve as an important resource for the intensively studied and harvested Galloanserae (ducks, upland game birds, and relatives).

Methods

Biological materials

A liver tissue sample was obtained from a male California quail (*C. californica brunnescens*) collected on 13 October 2020. This bird was captured at Cow Mountain Recreation Management Area, Lake Co., California (39.08005°N; 122.08910°W). The individual was collected with approval of the California Department of Fish and Wildlife (permit #: SCP-458) and the U.S. Fish and Wildlife Service (permit #: MB153526-0). The bird was captured and euthanized using methods approved by the University of California, Berkeley IACUC (AUP-2016-04-8665-1). A voucher specimen has been deposited at the Museum of Vertebrate Zoology, Berkeley, California (https://arctos.database.museum/guid/MVZ:Bird:193975).

PacBio HiFi library preparation

High molecular weight (HMW) genomic DNA (gDNA) was extracted from 45 mg of liver tissue using the Nanobind Tissue Big DNA kit as per the manufacturer's instructions (Pacific BioSciences—PacBio, Menlo Park, California). The DNA purity was estimated using absorbance ratios (260/280 = 1.80 and 260/230 = 2.06) on the NanoDrop ND-1000 spectrophotometer. The final DNA yield (17 µg) was quantified using the Quantus Fluorometer (QuantiFluor ONE dsDNA Dye assay; Promega, Madison, Wisconsin). The size distribution of the HMW DNA was estimated using the Femto Pulse system (Agilent, Santa Clara, California) and found that 71% of the fragments were 120 kb or more.

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 using Diagenode's Megaruptor 3 system (Diagenode, Belgium; Cat. B06010003). The sheared gDNA was concentrated using 0.45× of AMPure PB beads (PacBio; 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 for size selection using the BluePippin/PippinHT system (Sage Science, Beverly, Massachusetts; 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, California) 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, California) 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. An NGS library was generated using an NEB Ultra II DNA Library Prep kit (NEB, Ipswich, Massachusetts) with an Illumina compatible y-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 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 genome of the California quail following the CCGP assembly pipeline Version 4.0, as outlined in Table 1, which lists the tools and non-default parameters used in the assembly. The pipeline uses PacBio HiFi reads and Omni-C data to produce high-quality and highly contiguous genome assemblies minimizing manual curation. 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. 2022) 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 duplications, contig overlaps and repeats on the primary assembly with purge_dups (Guan *et al.* 2020) and transferred them to the alternate assembly. We aligned the Omni-C data to both assemblies following the Arima Genomics Mapping Pipeline (https://github.com/ArimaGenomics/mapping_pipeline) and then scaffolded both assemblies with SALSA (Ghurye et al. 2017, 2019).

We generated Omni-C contact maps for both assemblies by aligning the Omni-C data with BWA-MEM (Li 2013), identified ligation junctions, and generated Omni-C pairs using pairtools (Goloborodko *et al.* 2018). We generated a multiresolution Omni-C matrix with cooler (Abdennur and Mirny 2020) and balanced it with hicExplorer (Ramírez *et al.* 2018). We used HiGlass (Kerpedjiev *et al.* 2018) and the PretextSuite (https://github.com/wtsi-hpag/PretextView; https://github.com/wtsi-hpag/PretextMap; https://github.com/wtsi-hpag/PretextSnapshot) to visualize the contact maps and then checked the contact maps for major misassemblies. In detail, if in the proximity of a join that was made by the scaffolder we identified a strong off-diagonal signal and a lack of signal in the consecutive genomic region, we dissolved it by breaking the scaffolds at the coordinates of the join. After this process, no further manual joins were made. Some of the remaining gaps (joins generated by the scaffolder) were closed using the PacBio HiFi reads and YAGCloser (https://github.com/merlyescalona/yagcloser). Finally, we checked for contamination using the BlobToolKit Framework (Challis *et al.* 2020).

Mitochondrial genome assembly

We assembled the mitochondrial genome of *C. californica* from the PacBio HiFi reads using the reference-guided pipeline MitoHiFi (Allio *et al.* 2020; Uliano-Silva *et al.* 2021). The mitochondrial sequence of *C. douglasii* (https://www.ncbi.nlm.nih.gov/nuccore/MW574356.1; Kimball et al. 2021) 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 assembly assessment

We generated k-mer counts from the PacBio HiFi reads using meryl (https://github.com/marbl/meryl) and then used the k-mer database in GenomeScope2.0 (Ranallo-Benavidez *et al.* 2020) to estimate genome features including genome size, heterozygosity, and repeat content. We ran QUAST (Gurevich *et al.* 2013) to obtain general contiguity metrics, and used BUSCO (Manni *et al.* 2021) with the Aves ortholog database (aves_odb10, which contains 8,338 genes) to evaluate genome quality and functional completeness. 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. 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

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
Identification of haplotypic duplications	purge_dups (Custom cutoffs: 5,17,25,25,55,99)	1.2.5
Scaffolding		
Omni-C data alignment	Arima Genomics Mapping Pipeline	Commit 2e74ea4
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
Genome quality assessment		
Basic assembly metrics	QUAST (--est-ref-size)	5.0.2
Assembly completeness	BUSCO (-m geno, -l aves)	5.0.0
	Merqury	2020-01-29
Contamination screening		
Local alignment tool	BLAST+	2.1
General contamination screening	BlobToolKit	2.3.3
Synteny visualization		
Visualization of genome assembly consistency	JupiterPlot (ng = 80, m = 1,000,000)	1.0
Repeat analysis		
Identification of LTR elements	RepeatModeler (ltrstruct)	2
Annotation of TE diversity	RepeatMasker	4.1.2

Software citations are in the text.
[a]Options detailed for non-default parameters.

= Phred base accuracy QV (quality value); C = % genome represented by the first “n” scaffolds, following a known karyotype of 2n = 84 for *C. californica* (Bird Chromosome Database, Chromosome number data V3.0/2022, Degrandi *et al.* 2020). Quality metrics for the notation were calculated on the primary assembly.

To visualize higher-level synteny between quail scaffolds and chromosomes of the chicken genome, we generated a jupiter plot using the JupiterPlot pipeline (Chu 2018). We mapped the longest scaffolds representing 80% of the draft assembly to chicken chromosomes exceeding 1 Mb in length.

Repeat annotation

We performed de novo repeat annotation of the draft California quail reference assembly using the program RepeatModeler2 with the ltrstruct option selected to improve identification of LTR (Long Terminal Repeat) elements (Flynn *et al.* 2020). The resulting repeat library contained 496 elements and was used to annotate transposable element (TE) diversity in the California quail assembly using RepeatMasker (Smit *et al.* 2013–2015). To assess temporal patterns of TE activity in the quail genome, we used the calcDivergenceFromAlign.pl script in the RepeatMasker package to estimate the Kimura 2-parameter (K2P) distance of each TE element from the consensus sequence. K2P distances were used to generate barplots for the LTR, SINE, LINE, and DNA classes of TE elements found in the genome.

Results

Sequencing data

The Omni-C and PacBio HiFi sequencing libraries generated 132.2 million read pairs and 3.1 million reads, respectively. The latter yielded ~45-fold coverage based on the Genomescope2.0 genome size estimation of 1.09 Gb (N50

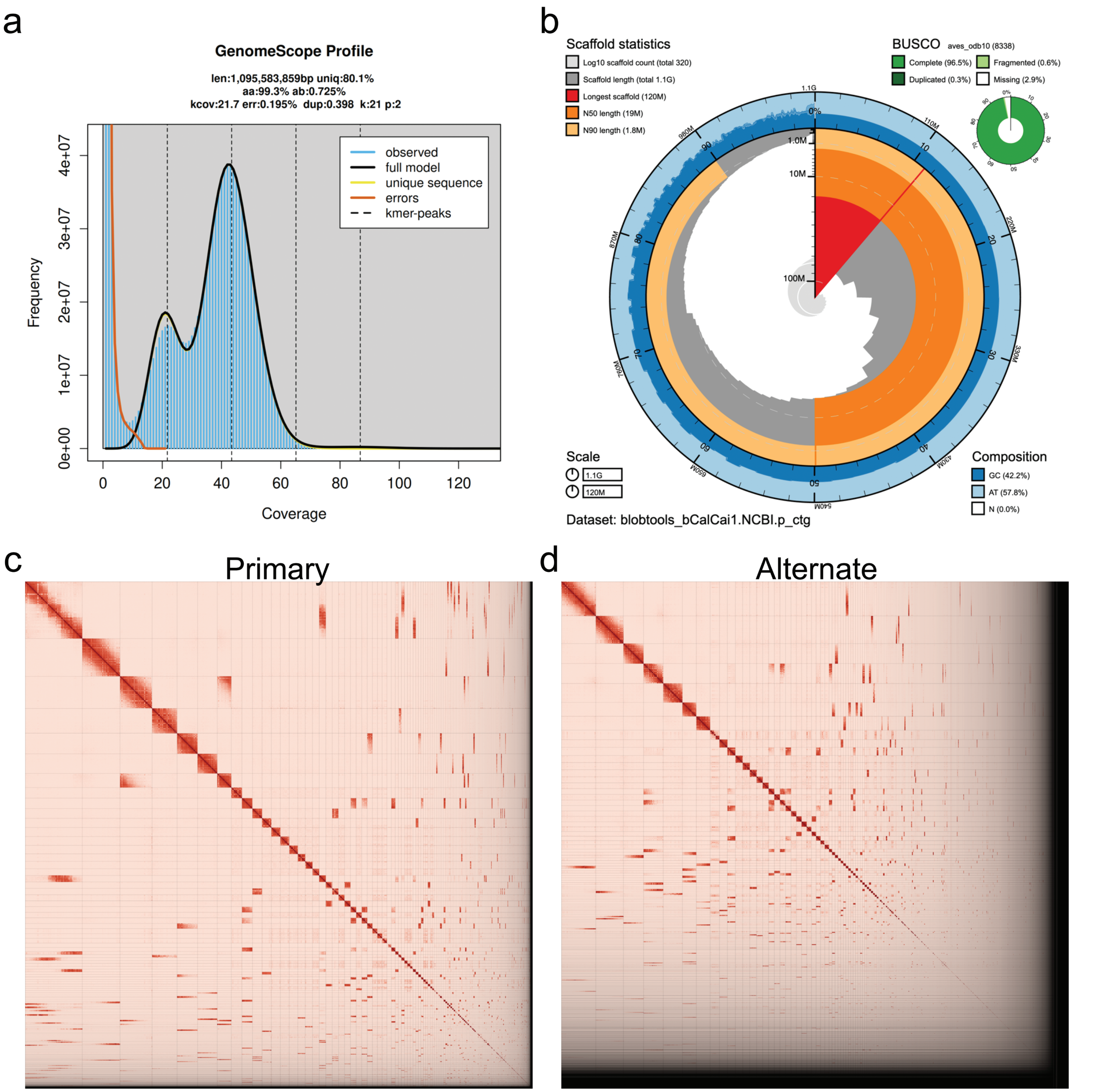

Fig. 2. Visual overview of genome assembly metrics. a) K-mer spectra output generated from PacBio HiFi data without adapters using GenomeScope2.0. The bimodal pattern observed corresponds to a diploid genome and the k-mer profile matches that of low (<1%) heterozygosity. K-mers covered at lower coverage 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 *Callipepla californica* primary assembly (bCalCai1). 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 such regions.

read length of 15,835 bp, minimum read length 44 bp; mean read length 15,667 bp, maximum read length 50,092 bp). Based on the PacBio HiFi reads, we estimated 0.195% sequencing error rate and 0.7% heterozygosity rate. The k-mer spectrum, also based on the PacBio HiFi reads, shows a bimodal distribution with major peaks at ~22- and ~45-fold coverage, where peaks correspond to heterozygous and homozygous states of a diploid species, respectively (Fig. 2a). The distribution presented in this k-mer spectrum supports that of a low heterozygosity profile.

Table 2. Metrics for the primary (left column) and alternate (right column) assemblies of the California quail (*Callipepla californica*) genome.

BioProjects & Vouchers	CCGP NCBI BioProject			PRJNA720569			
	Genera NCBI BioProject			PRJNA766282			
	Species NCBI BioProject			PRJNA777150			
	NCBI BioSample			SAMN26894057			
	Specimen voucher			MVZ:Bird:193975			
	NCBI Genome accessions			Primary		Alternate	
	Assembly accession			JALIRH000000000		JALIRI000000000	
	Genome sequences			GCA_023055505.1		GCA_023055725.1	
Genome Sequence	PacBio HiFi reads	Run		1 PACBIO_SMRT (Sequel II) run: 3.2M spots, 49.6G bases, 35.9 Gb			
		Accession		SRX15651218			
	Omni-C Illumina reads	Run		2 ILLUMINA (Illumina NovaSeq 6000) runs: 132.2M spots, 39.9G bases, 12.8 Gb			
		Accession		SRX15651219, SRX15651220			
Genome Assembly Quality Metrics	Assembly identifier (quality code[a])			bCalCai1(6.7.P6.Q63.C78)			
	HiFi read coverage[b]			45.28×			
				Primary		Alternate	
	Number of contigs			608		1,855	
	Contig N50 (bp)			5,506,990		3,481,594	
	Contig NG50[b]			5,478,314		5,669,995	
	Longest contigs			25,292,651		30,636,471	
	Number of scaffolds			321		1,607	
	Scaffold N50			19,408,658		12,585,889	
	Scaffold NG50[b]			18,906,995		16,621,059	
	Largest scaffold			121,926,015		94,958,808	
	Size of final assembly			1,085,347,722		1,399,796,956	
	Phased block NG50[b]			5,365,876		5,773,240	
	Gaps per Gbp (# Gaps)			264 (287)		177 (248)	
	Indel QV (frameshift)			42.85789529		42.86055461	
	Base pair QV			63.6319		62.6168	
				Full assembly = 63.0313			
	K-mer completeness			89.7222		90.4135	
				Full assembly = 99.4094			
	BUSCO completeness (aves_odb10), *n* = 8,338		C	S	D	F	M
		P[c]	96.50%	96.20%	0.30%	0.60%	2.90%
		A[c]	96.70%	96.20%	0.50%	0.60%	2.70%
	Organelles		1 Partial mitochondrial sequence			JALIRH010000321.1	

(a) 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 = 82 following a known karyotype of 2n =84 for C. californica (Bird Chromosome Database, Chromosome number data V3.0/2022, Degrandi *et al.* 2020). Quality code for all the assembly denoted by primary assembly (bCalCai1.0.p).
(b) Read coverage and NGx statistics have been calculated based on the estimated genome size of 1.09 Gb.
(c) (P)rimary and (A)lternate assembly values.
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/data base. bp, base pairs.

Assembly metrics

The final assembly (bCalCai1) consists of 2 pseudo haplotypes, here tagged as primary and alternate, with the genome size of the primary assembly similar to the estimated genome size from Genomescope2.0 (Fig. 2a). The primary assembly consists of 321 scaffolds spanning 1.08 Gb with contig N50 of 5.5 Mb, scaffold N50 of 19.4 Mb, largest contig N50 of 25.2 Mb, and largest scaffold N50 of 121.9 Mb. On the other hand, the alternate assembly consists of 1,855 scaffolds spanning 1.39 Gb with contig N50 of 3.4 Mb, scaffold N50 of 12.5 Mb, largest contig of 30.6 Mb, and largest scaffold of 94.9 Mb. Detailed assembly statistics are reported in Table 2 and graphical representation for the primary assembly is shown in Fig. 2b (see Supplementary Fig. S1 for the alternate assembly). Hi-C contact maps for the primary and alternate assemblies are shown in Fig. 2c and d, respectively. The primary assembly has a BUSCO completeness score of 96.5% using the Aves gene set, a per-base quality (QV) of 63.63, a k-mer completeness of 89.72, and a frameshift indel QV of 42.85; the alternate assembly has a BUSCO completeness score of 96.7% using the Aves gene set, a per-base quality (QV) of 62.61, a k-mer completeness of 90.41, and a frameshift indel QV of 42.86.

We identified 16 misassemblies, 11 on the primary and 5 on the alternate, and broke the corresponding joins made by SALSA. We were able to close 16 gaps, 10 on the primary and 6 on the alternate. Finally, we removed 2 contigs, one per assembly, corresponding to mitochondrial contaminants. No further sequences were removed. We have deposited both

Table 3. A comparison of assembly metrics and BUSCO scores for the California quail, 3 other Odontophoridae quail genomes (scaled quail, northern bobwhite, marbled wood-quail) and the chicken assembly.

	California quail (*Callipepla californica*)	Scaled quail (*Callipepla squamata*)	Northern bobwhite (*Colinus virginianus*)	Marbled wood-quail (*Odontophorus gujanensis*)	Chicken (*Gallus gallus*)
Assembly	GCA_023055505.1	GCA_002218305.1	GCA_008692595.2	GCA_013399175.1	GCA_000002315.5
Total length	1,085,347,722	1,045,281,893	948,540,911	977,719,362	1,065,348,650
# contigs	608	59,818	39,169	52,876	1,402
Contig N50	5,506,990	154,132	106,817	74,578	17,655,422
Contig L50	53	1,950	2,337	3,724	19
# scaffolds	321	34,302	6,035	26,662	524
Scaffold N50	19,408,658	1,035,259	66,825,924	343,789	20,785,086
Scaffold L50	12	278	5	781	12
BUSCO results (%):					
Complete	96.5	94.2	90.8	93.5	—
Single copy	96.2	92.8	89.8	92.5	—
Duplicate	0.3	1.4	0.9	1	—
Fragmented	0.6	3.6	3.5	3.7	—
Missing	2.0	2.2	5.8	2.8	—

BUSCO scores for the California quail genome were based on the aves_odb10 database, scores for the other quail genomes were based on earlier aves databases. Data for scaled quail and marbled wood-quail from Feng *et al.* (2020), northern bobwhite from Salter *et al.* (2019), and chicken from Warren *et al.* (2017).

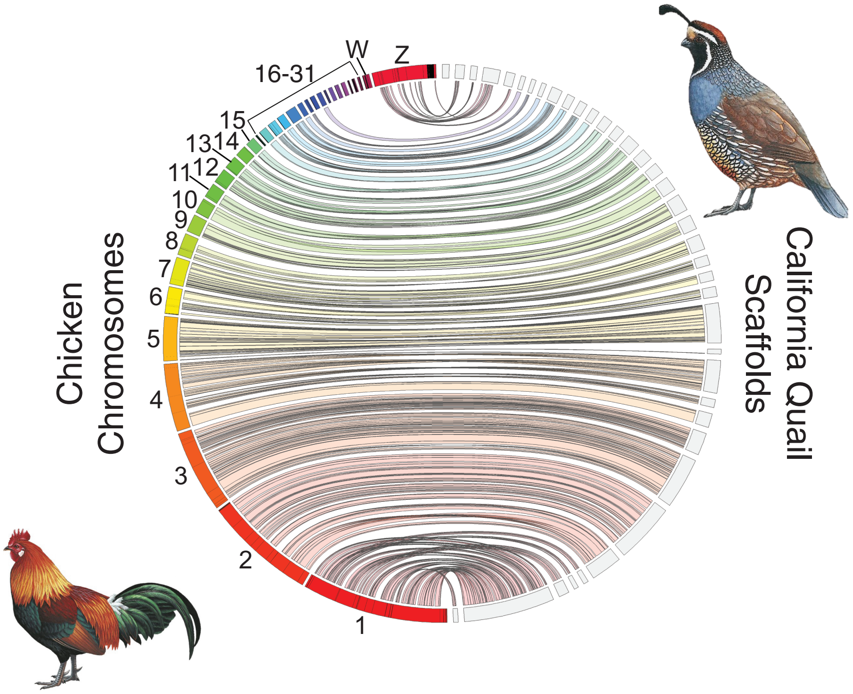

Fig. 3. Jupiter plot comparing higher-level synteny and completeness between the chicken (*Gallus gallus*) genome (GRCg6a) and the California quail draft assembly. Chicken chromosomes are on the (colored) and quail scaffolds are on the right (light gray). Twists represent reversed orientation of scaffolds between assemblies. Bird illustrations reproduced from https://birdsoftheworld.org with permission from Lynx Edicions.

assemblies on NCBI (see Table 2 and Data Availability for details).

All metrics for the California quail assembly exceed those reported for other Odontophoridae quail genomes with the exception of a higher scaffold N50 in the most recent northern bobwhite genome (Table 3). A Jupiter plot showed quail scaffolds mapping to most chromosomes of the chicken genome with little evidence for major inversions or translocations (Fig. 3). The Jupiter plot also shows 2 large scaffolds (Scaffold 2: 80 Mb; Scaffold 6: 42 Mb) mapping to chicken chromosome 2. This likely corresponds to a previously reported fission of chicken chromosome 2 into chromosomes 3 and 6 of the California quail (Shibusawa *et al.* 2004).

Table 4. Percentage of each genome spanned by different classes of repeats.

	Number elements	Length (bp)	Percentage of genome (%)
Retroelements	273,188	135,663,026	12.5
SINEs	3,230	414,480	0.04
LINEs	203,697	74,892,484	6.9
LTR elements	66,261	60,356,062	5.56
DNA transposons	28,034	8,366,846	0.77
Unclassified	60,686	34,668,306	3.19
Total interspersed repeats		178,698,178	16.46
Other repeats			
Small RNA	885	106,771	0.01
Satellites	469	292,684	0.03
Simple repeats	299,705	24,793,328	2.28
Low complexity	54,030	3,734,179	0.34

Estimates of each class of repeat region identified within RepeatMasker using the quail TE libraries generated de novo with RepeatModeler2.

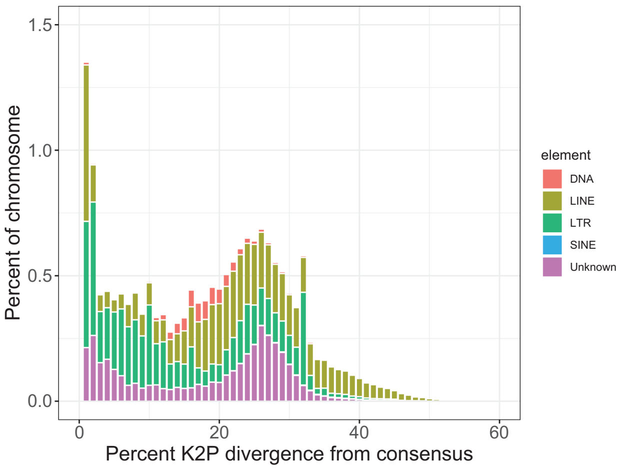

Fig. 4. TE landscape for the California quail genome. Percent divergence on the *x* axis was calculated as the percent Kimura 2-parameter (K2P) distance with excluding CpG sites. The abundance of TEs in each percent divergence bin was normalized as percentage of the genome length on the *y* axis.

Mitochondrial assembly

We assembled a mitochondrial genome for *C. californica* with MitoHiFi. The final mitochondrial sequence has a size of 16,706 bp. The base composition is A= 30.52%, C = 31.99, G = 13.26%, T = 24.23, and the sequence consists of 22 unique transfer RNAs and 13 protein-coding genes.

Repeat annotation

19.12% of the California quail genome was masked by RepeatMasker (Table 4). This included 178.7 Mb (16.46%) of the genome masked by interspersed repeats and 28.8 Mb (2.66%) masked by satellites and other simple repeats. Endogenous retroviral elements (5.25%) and LINE elements of the CR1 family (6.90%) were the most abundant TEs annotated in the quail genome. DNA transposons and SINE elements were each found in <1% of the genome. LINE elements exhibited a pattern of greater divergence from the consensus sequence (peak between 20% and 30% K2P divergence), indicating that these elements likely stem from a more ancient proliferation. In contrast, LTR elements exhibit a pattern of lower divergence from the consensus i.e. indicative of more recent activity (Fig. 4).

Discussion

We generated a highly contiguous genome for the California quail (*C. californica*). The contig N50 for this assembly exceeded 91% of all avian genome assemblies generated to date, and scaffold N50 was in the 86th percentile for all avian genomes (Bravo *et al.* 2021). Compared with 3 other Odontophoridae quail genomes, the California quail assembly exhibited the highest BUSCO completeness (96.5% of 8,338 orthologs present) and was the only assembly with a contig N50 exceeding 1 Mb in length. A high-quality northern bobwhite (*Colinus virginianus*) assembly had a much higher scaffold N50 of 66.8 Mb but a contig N50 of 0.11 Mb (Salter *et al.* 2019). Similarly, the California quail genome contains only 2.64 Ns per 100 kb, whereas the bobwhite genome has 1,363.4 Ns per 100 kb. These differences underscore the power of PacBio long reads for generating a more contiguous and complete genome.

TE sequence spans ~16.5% of the California quail genome. This is higher than estimates from previous quail assemblies based on short-read sequencing data, which included on average 9.2% TE content (Feng *et al.* 2020). However, the higher levels of TE content found in the quail genome more closely matches levels observed in other highly contiguous genomes generated from other lineages within Galloanserae (ducks and chickens). The galGal5 assembly for the domestic chicken (*G. gallus*; family Phasianidae; order Galliformes) found 16.5% of the genome to be repeats (Warren *et al.* 2017), whereas both wild and domestic mallard (*A. platyrhynchos*; family Anatidae; order Anseriformes) genomes were found to contain ~17% TE content (Zhu *et al.* 2021). In California quail, the increase in detectable TE content was primarily driven by newly discovered LTR element sequences. The ability to detect a greater abundance of LTR elements in long-read assemblies has been shown in a number of species, including the duck genome. LTR elements tend to be longer than other TE sequences and have been shown to be a major cause of gaps in avian assemblies

(Peona *et al*. 2021). The first comparisons of TE diversity among the genomes of chicken, turkey, and zebra finch suggested a much greater abundance of LTRs in the zebra finch relative to galliform genomes (Warren *et al*. 2010). The California quail genome joins a growing number of highly contiguous genomes from within Galloanserae that together point to a much greater abundance of LTR elements than previously appreciated.

A high-quality reference genome for California quail will be an important resource for research and conservation of this popular upland game bird. First, California quail is a bird of significant recreational interest throughout its native and introduced distribution. For example, an estimated 315,268 quail were harvested by over 35,000 hunters in California during the 2018 to 2019 season (Miller and Meshriy 2019). However, management units are geographically broad and do not currently incorporate population genetic data that can inform patterns of population structure or genetic diversity. Second, population genetic data aligned to this reference genome will be important for determining the distinctiveness of the Catalina Island subspecies (*C. c. catalinensis*) and for testing speculations about human transport to the Channel Islands in the last 12,000 yr (Collins 2008). Third, this will facilitate research on speciation and hybridization genomics between Gambel's and California quail. Of particular interest will be the impact of differences in chromosome number between the 2 species (Gambel's: 2n = 80 chromosomes, California: 2n = 84; Degrandi *et al*. 2020) on hybrid zone dynamics. Finally, a major priority for improved quail management and conservation will be understanding local adaptation across rainfall gradients and the capacity of quail populations to adapt to future changes in rainfall patterns across the west (Zornes and Bishop 2009). The impacts of annual rainfall on reproductive success have been well established in quail, with aridland populations responding positively to increased annual rainfall but more mesic-adapted populations responding negatively (Leopold 1977). Generation of a high-quality reference genome is a critical step toward understanding the genetic basis of local adaptation across precipitation gradients in this species.

Supplementary material

Supplementary material is 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].

Acknowledgments

We would like to thank Joshua Ho for assistance with tissue subsampling, Jackie Childers for contributing photos of both male and female California quail, and Zachary Harlow for providing access to the Blue Oak Ranch Reserve, Santa Clara County, California. PacBio Sequel II library prep and sequencing were carried out at the DNA Technologies and Expression Analysis Cores at the UC Davis Genome Center, supported by NIH Shared Instrumentation Grant 1S10OD010786-01. Deep sequencing of Omni-C libraries used the Novaseq S4 sequencing platforms at the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 OD018174 Instrumentation Grant. We thank the staff at the UC Davis DNA Technologies and Expression Analysis Cores and the UC Santa Cruz Paleogenomics Laboratory for their diligence and dedication to generating high-quality sequence data.

Data availability

Data generated for this study are available under the NCBI BioProject PRJNA720569. Raw sequencing data for the California quail (NCBI BioSample SAMN26894057) are deposited in the NCBI Short Read Archive (SRA) under SRS13350135. The transposable element library used to mask repeats has been submitted to dryad. Assembly scripts and other data used for the analyses are deposited in the following GitHub repository: www.github.com/ccgproject/ccgp_assembly.

References

Abdennur N, Mirny LA. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics*. 2020;36(1):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 Resour*. 2020;20(4):892–905.

Bravo GA, Schmitt CJ, Edwards SV. What have we learned from the first 500 avian genomes? *Annu Rev Ecol Evol Syst*. 2021;52(1):611–639. doi:10.1146/annurev-ecolsys-012121-085928

Calkins JD, Gee JM, Hagelin JC, Lott DF. California quail (Callipepla californica), version 1.0. In: Poole AF, editor. *Birds of the world*. Ithaca (NY): Cornell Lab of Ornithology; 2014.

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

Challis R, Richards E, Rajan J, Cochrane G, Blaxter M. BlobToolKit—interactive quality assessment of genome assemblies. *G3*. 2020;10(4):1361–1374.

Cheng H, Jarvis ED, Fedrigo O, Koepfli KP, Urban L, Gemmell NJ, Li H. Haplotype-resolved assembly of diploid genomes without parental data. *Nature Biotechnology*. 2022;40(9):1332–1335. https://doi.org/10.1038/s41587-022-01261-x

Chu J. Jupiter Plot: a Circos-based tool to visualize genome assembly consistency (1.0). *Zenodo*. 2018. doi:10.5281/zenodo.1241235

Collins PW. Catalina California Quail (*Callipepla californica catalinensis*). In: Shuford WD, Gardali T, editors. *California bird species of special concern: a ranked assessment of species, subspecies, and distinct populations of birds of immediate conservation concern in California. Studies of Western Birds 1*. Camarillo (CA), Sacramento: Western Field Ornithologists, California Department of Fish and Game; 2008.

Degrandi TM, Barcellos SA, Costa AL, Garnero ADV, Hass I, Gunski RJ. Introducing the bird chromosome database: an overview of cytogenetic studies in birds. *Cytogenet Genome Res*. 2020;160(4):199–205. doi:10.1159/000507768

Feng S, Stiller J, Deng Y, Armstrong J, Fang Q, Reeve AH, Xie D, Chen G, Guo C, Faircloth BC, et al. Dense sampling of bird diversity increases power of comparative genomics. *Nature*. 2020;587(7833):252–257.

Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci USA*. 2020;117(17):9451–9457.

Ghurye J, Pop M, Koren S, Bickhart D, Chin C-S. Scaffolding of long read assemblies using long range contact information. *BMC Genomics*. 2017;18(1):527.

Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, Koren S. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS Comput Biol*. 2019;15(8):e1007273.

Goloborodko A, Abdennur N, Venev S, Brandao HB, Fudenberg G. *Zenodo*. 2018. doi:10.5281/zenodo.1490831

Grinnell JG, Bryant HC, Storer TI. *The game birds of California*. Berkeley (CA): University of California Press; 1918.

Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*. 2020;36(9):2896–2898.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. 2013;29(8):1072–1075.

Hosner PA, Braun EL, Kimball RT. Land connectivity changes and global cooling shaped the colonization history and diversification of New World quail (Aves: Galliformes: Odontophoridae). *J Biogeogr*. 2015;42(10):1883–1895.

Iknayan KJ, Wheeler MM, Safran SM, Young JS, Spotswood EN. What makes urban parks good for California quail? Evaluating park suitability, species persistence, and the potential for reintroduction into a large urban national park. *J Appl Ecol*. 2021;59(1):199–209.

Kerpedjiev P, Abdennur N, Lekschas F, McCallum C, Dinkla K, Strobelt H, Gehlenborg N. HiGlass: web-based visual exploration and analysis of genome interaction maps. *Genome Biol*. 2018;19(1):125.

Kimball RT, Guido M, Hosner PA, Braun EL. When good mitochondria go bad: cyto-nuclear discordance in landfowl (Aves: Galliformes). *Gene*. 2021 July;801:145841.

Korlach J, Gedman G, Kingan SB, Chin C-S, Howard JT, Audet J-N, 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(10):1–16.

Leopold AS. *The California quail*. Berkeley (CA): University of California Press; 1977.

Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, 2013, https://doi.org/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, *Molecular Biology and Evolution*, 2021;38(10):4647–4654. https://doi.org/10.1093/molbev/msab199

Miller K, Meshriy M. *Resident Upland Game Bird and Small Game Mammal Harvest Survey 2018–2019*. Sacramento: California Department of Fish and Wildlife; 2019.

Peona V, Blom MPK, Xu L, Burri R, Sullivan S, Bunikis I, Liachko I, Haryoko T, Jonsson KA, Zhou Q, et al. Identifying the causes and consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise. *Mol Ecol Resour*. 2021;21(1):263–286.

Ramírez F, Bhardwaj V, Arrigoni L, Lam KC, Grüning BA, Villaveces J, Haberman B, Akhtar A, Manke T. High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat Commun*. 2018;9(1):1–9. doi:10.1038/s41467-017-02525-w

Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun*. 2020;11(1):1432.

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(7856):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):245.

Salter JF, Hosner PA, Tsai WLE, McCormack JE, Braun EL, Kimball RT, Faircloth BC. Historical specimens and the limits of subspecies phylogenomics in the New World quails (Odontophoridae). *Mol Phylogenet Evol*. 2022 July;175:107559.

Salter JF, Johnson O, Stafford NJ, Herrin WF, Schilling D, Cedotal C, Faircloth BC. A highly contiguous reference genome for northern bobwhite (*Colinus virginianus*). *G3*. 2019;9(12):3929–3932.

Shaffer HB, Toffelmier E, Corbett-Detig RB, Escalona M, Erickson B, Fiedler P, Wang IJ. Landscape genomics to enable conservation actions: the California Conservation Genomics Project. *J Hered*. 2022;113(6):577–588. doi:10.1093/jhered/esac020

Shibusawa M, Nishida-Umehara C, Tsudzuki M, Masabanda J, Griffin DK, Matsuda Y. A comparative karyological study of the blue-breasted quail (*Coturnix chinensis*, Phasianidae) and California quail (*Callipepla californica*, Odontophoridae). *Cytogenet Genome Res*. 2004;106(1):82–90.

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):157.

Smit AFA, Hubley R, Green P. RepeatMasker Open-4.0. 2013–2015. http://www.repeatmasker.org. Accessed 15 October 2022.

Uliano-Silva M, Ferreira Nunes JG, Krasheninnikova K, McCarthy SA. marcelauliano/MitoHiFi: mitohifi_v2.0 (v2.0). *Zenodo*. 2021. doi:10.5281/zenodo.5205678

Warren WC, Clayton DF, Ellegren H, Arnold AP, Hillier LW, Künstner A, Searle S, White S, Vilella AJ, Fairley S, et al. The genome of a songbird. *Nature*. 2010;464(7289):757–762.

Warren WC, Hillier LDW, Tomlinson C, Minx P, Kremitzki M, Graves T, Markovic C, Bouk N, Pruitt KD, Thibaud-Nissen F, et al. A new chicken genome assembly provides insight into avian genome structure. *G3*. 2017;7(1):109–117.

Winkler DW, Billerman SM, Lovette IJ. New World Quail (*Odontophoridae*), version 1.0. In: Billerman SM, Keeney BK, Rodewald PG, Schulenberg TS, editors. *Birds of the world*. Ithaca (NY): Cornell Lab of Ornithology; 2020.

Zhu F, Yin ZT, Wang Z, Smith J, Zhang F, Martin F, Ogeh D, Hincke M, Lin FB, Burt DW, et al. Three chromosome-level duck genome assemblies provide insights into genomic variation during domestication. *Nat Commun*. 2021;12(1):1–11.

Zink RM, Lott DF, Anderson DW. Genetic variation, population structure, and evolution of California quail. *Condor*. 1987;89(2):395–405.

Zonana DM, Gee JM, Bridge ES, Breed MD, Doak DF. Assessing behavioral associations in a hybrid zone through social network analysis: complex assortative behaviors structure associations in a hybrid quail population. *Am Nat*. 2019;193(6):852–865.

Zornes M, Bishop RA. In: Williamson SJ, editor. *Western Quail Conservation Plan*. Cabot (VT): Wildlife Management Institute; 2009. p. 92.