Title

A highly contiguous genome assembly for the Yellow Warbler (Setophaga petechia)

Journal

Journal of Heredity, 115(3)

ISSN

0022-1503

Authors

Tsai, Whitney LE

Escalona, Merly

Garrett, Kimball L

et al.

Publication Date

2024-05-09

DOI

10.1093/jhered/esae008

Peer reviewed

Journal of Heredity, 2024, **115**, 317–325
https://doi.org/10.1093/jhered/esae008
Advance access publication 24 February 2024
Genome Resources

American Genetic Association

OXFORD

Genome Resources

A highly contiguous genome assembly for the Yellow Warbler (*Setophaga petechia*)

Whitney L.E. Tsai[1,2,*], Merly Escalona[3], Kimball L. Garrett[4], Ryan S. Terrill[2], Ruta Sahasrabudhe[5], Oanh Nguyen[5], Eric Beraut[6], William Seligmann[6], Colin W. Fairbairn[6], Ryan J. Harrigan[1], John E. McCormack[2], Michael E. Alfaro[1], Thomas B. Smith[1] and Rachael A. Bay[7]

[1]Department of Ecology and Evolutionary Biology, University of California, Los Angeles, CA 90095, United States,
[2]Moore Laboratory of Zoology, Biology Department, Occidental College, Los Angeles, CA 90041, United States,
[3]Department of Biomolecular Engineering, University of California, Santa Cruz, CA 95064, United States,
[4]Ornithology Department, Natural History Museum of Los Angeles County, Los Angeles, CA 90007, United States,
[5]DNA Technologies and Expression Analysis Core Laboratory, Genome Center, University of California, Davis, CA 95616, United States,
[6]Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, CA 95064, United States,
[7]Department of Evolution and Ecology, University of California, Davis, CA 95616, United States

[*]Address correspondence to W. L. E. Tsai at the address above, or email: whitney.le.tsai@gmail.com

Corresponding Editor: Alexander Suh

Abstract

The Yellow Warbler (*Setophaga petechia*) is a small songbird in the wood-warbler family (Parulidae) that exhibits phenotypic and ecological differences across a widespread distribution and is important to California's riparian habitat conservation. Here, we present a high-quality de novo genome assembly of a vouchered female Yellow Warbler from southern California. Using HiFi long-read and Omni-C proximity sequencing technologies, we generated a 1.22 Gb assembly including 687 scaffolds with a contig N50 of 6.80 Mb, scaffold N50 of 21.18 Mb, and a BUSCO completeness score of 96.0%. This highly contiguous genome assembly provides an essential resource for understanding the history of gene flow, divergence, and local adaptation in Yellow Warblers and can inform conservation management of this charismatic bird species.

Key words: California Conservation Genomics Project, Parulidae

Introduction

The Yellow Warbler (*Setophaga petechia*) is a widespread songbird species distributed from Alaska to northern South America (Fig. 1). The species complex comprises up to 43 subspecies in four distinct subspecies groups that display notable diversity in phenotype and ecology across their range (Browning 1994; Klein and Brown 1994; Wilson and Holberton 2004; Salgado-Ortiz et al. 2008). This phenotypic diversity and the presence of both migratory and resident populations have encouraged investigation into the history of adaptation, divergence, and gene flow in this species (Gibbs et al. 2000; Milot et al. 2000; Chaves et al. 2012; Chavarria-Pizarro et al. 2019; Machkour-M'Rabet et al. 2023). Additionally, as a widespread migratory bird species, the Yellow Warbler inhabits variable environmental conditions across its range, allowing for the investigation into the influence of climate on geographic variation and genomic capacity to adapt to climate change (Bay et al. 2018; Chen et al. 2022; DeSaix et al. 2022).

In California, Yellow Warblers are listed as a Species of Special Concern (Shuford et al. 2008) and have experienced notable declines over the last 50 years (Sauer et al. 2014). Previous genomic work indicates that the inability to adapt to climate change may play a role in population declines in California (Bay et al. 2018). California wetlands and riparian corridors are crucial stopover and breeding habitats for Yellow Warblers and other species of migratory birds. In the last century, 90% to 95% of historic wetland and riparian habitats have been lost, and those that remain are threatened by development and climate change (Dahl 1990; Krueper 1996; Poff et al. 2012). As indicators of healthy riparian habitat, understanding how California Yellow Warbler populations adapt to dramatic changes in their environment will inform conservation action and help mitigate habitat loss in other vulnerable and threatened riparian species, like the California Red-legged Frog (*Rana draytonii*), the Riparian Brush Rabbit (*Sylvilagus bachmani riparius*), and the Valley Elderberry Longhorn Beetle (*Desmocerus californicus dimorphus*) (Collinge et al. 2001; Davidson et al. 2001; Heath and Ballard 2003; Phillips et al. 2005).

The evolutionary and conservation genomics studies needed to address these questions increasingly rely on low-coverage,

Received December 22, 2023; Accepted February 16, 2024

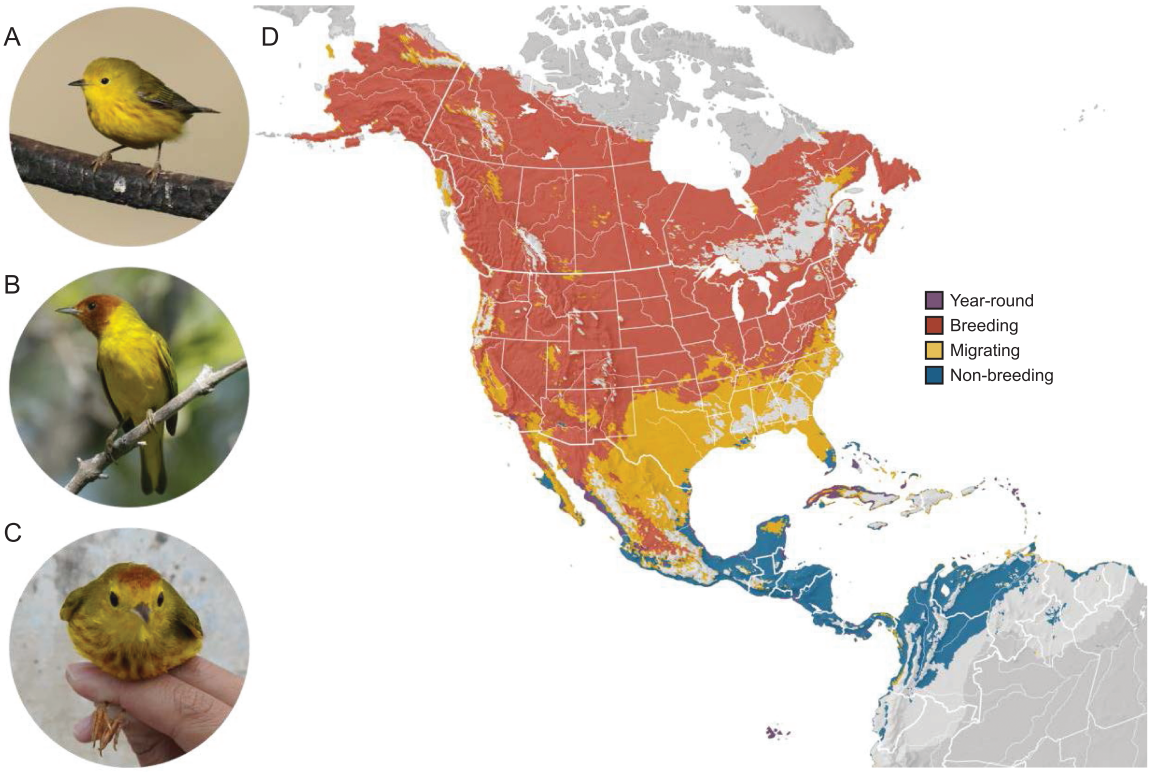

Fig. 1. Geographic variation and distribution of Yellow Warblers (*Setophaga petechia*). A) The Northern (*aestiva*) group includes migratory subspecies with chestnut streaking on the breast. Northern subspecies breed in North America and winter in Central and northern South America. Photo taken by R. S. Terrill at Piute Ponds, Los Angeles, CA, USA. B) The Mangrove (*erithachorides*) group includes resident subspecies with a characteristic chestnut head. Mangrove subspecies inhabit mangroves along the coasts of Central and northern South America year-round. Photo taken by R. S. Terrill on Isla Holbox, Quintana Roo, MX. C) The Galapagos (*aureola*) and Golden (*petechia*) subspecies groups include resident subspecies with a chestnut cap and thick breast streaking except for *S. p. ruficapilla* from Martinique which exhibits the Mangrove phenotype. Populations of the Galapagos subspecies are found on the Galapagos Islands and Cocos Island off Costa Rica and Golden subspecies are found on the islands of the Caribbean. Photo taken by W. L. E. Tsai on Isla Cozumel, Quintana Roo, MX. D) Map of species distributional abundance (Fink et al. 2022). Shaded colors indicate seasonal shifts in distributions: year-round (purple), breeding (red), migrating (yellow), and non-breeding (blue).

whole genome sequencing (WGS), which requires a high-quality reference genome for alignment. Reference genome assemblies provide a map of the structural features and organization of the genome and the choice of reference genome assembly for WGS studies can impact evolutionary inferences like demographic history and genetic diversity (Gopalakrishnan et al. 2017). Currently, there are four genome assemblies generated with short-read sequencing technology for the genus *Setophaga*. There is one Yellow-rumped Warbler (*S. coronata*) chromosome-level assembly (Toews et al. 2016), two Kirtland's Warbler (*S. kirtlandii*) scaffold-level assemblies (Feng et al. 2020), and the existing draft genome assembly for Yellow Warbler has a length of 1.26 Gb, a total of 18,414 scaffolds, and a scaffold N50 491.7 kb (Bay et al. 2018). The use of an interspecific reference genome assembly can lead to many errors and biases, including lower mapping ability (especially in regions with higher evolutionary rates) and inaccurate gene order (Prasad et al. 2022). The high number and relatively short scaffold length of the existing Yellow Warbler genome assembly could hinder the identification of structural variants often maintained between and within species and are important in adaptive evolution, speciation, and generating morphological diversity (Lamichhaney et al. 2016; Wellenreuther and Bernatchez 2018; Mérot et al. 2020). Additionally, reference genome assemblies generated solely from short-read sequencing technology fail to resolve lengths and placement of repeat regions, such as transposable elements or telomeres, leading to gaps in avian genome assemblies (Peona et al. 2021). This highlights the need for a high-quality, species-specific reference genome for WGS studies.

Here, we present a new genome assembly for the Yellow Warbler generated as part of the California Conservation Genomics Project (CCGP) consortium (Shaffer et al. 2022). We used high-molecular-weight (HMW) genomic DNA (gDNA) extracted from a vouchered, female bird collected in California and leveraged Pacific Biosciences (PacBio) HiFi long-read and Dovetail Genomics Omni-C proximity sequencing technologies. This produced a high-quality genome assembly that will allow us to better understand evolutionary processes like phenotypic variation and migration and conduct conservation genomics studies to inform conservation initiatives.

Methods

Biological materials

We sampled heart, liver, muscle, and other tissues from a female Yellow Warbler collected using mist nets near Stephen Sorensen Park (34.60549°N, 117.8306°W) in Los Angeles County, California on 25 September 2020. This migrant Yellow Warbler can presumably be assigned to *Setophaga petechia brewsteri* based on collection date and locality (Browning 1994) and was

collected with approval from the following entities: California Department of Fish and Wildlife Scientific Collecting Permit (#SC-000939), US Fish and Wildlife Services Scientific Collecting Permit (MB708062-0), and US Geological Survey Banding Permit (22804-B). Tissue samples were retrieved and flash-frozen in liquid nitrogen, and the first muscle tissues were frozen within 2 min of specimen collection. A voucher specimen and tissue are deposited at the Natural History Museum of Los Angeles (LACM Bird #122168, KLG4550, LAF9440). Additional tissues for this individual are housed in the CCGP tissue repository at the University of California, Los Angeles under identification YEWA_CCGP3.

Nucleic acid extraction, library preparation, and sequencing

We extracted HMW gDNA from 30 mg of flash-frozen heart tissue. We homogenized the tissue by grinding it in a mortar and pestle in liquid nitrogen. We lysed the homogenized tissue at room temperature overnight with 2 ml of lysis buffer containing 100 mM NaCl, 10 mM Tris-HCl pH 8.0, 25 mM EDTA, 0.5% (w/v) SDS, and 100 µg/ml Proteinase K. We treated the lysate with 20 µg/ml RNAse at 37 °C for 30 min. We cleaned the lysate with equal volumes of phenol/chloroform using phase lock gels (Quantabio, MA; Cat # 2302830). We precipitated the DNA from the cleaned lysate by adding 0.4× volume of 5M ammonium acetate and 3× volume of ice-cold ethanol. We washed the pellet twice with 70% ethanol and resuspended it in elution buffer (10 mM Tris, pH 8.0). We measured DNA purity using absorbance ratios (260/280 = 1.87 and 260/230 = 2.29) using a NanoDrop ND-1000 spectrophotometer. We quantified DNA yield (30 µg) using a Qubit 2.0 Fluorometer (Thermo Fisher Scientific, MA). We verified HMW gDNA integrity on a Femto pulse system (Agilent Technologies, CA), where 80% of the DNA was found in fragments above 120 kb.

According to the manufacturer's instructions, we constructed the HiFi Single Molecule, Real-Time (SMRT) library using SMRTbell Express Template Prep Kit v2.0 (PacBio, CA; Cat. #100-938-900). We sheared HMW gDNA to a target DNA size distribution between 15 and 20 kb and concentrated it using 0.45× of AMPure PB beads (PacBio; Cat. #100-265-900). We performed the enzymatic incubations as follows: removal of single-strand overhangs at 37 °C for 15 min, DNA damage repair at 37 °C for 30 min, end repair at 20 °C for 10 min, A-tailing at 65 °C for 30 min, ligation of overhang adapter v3 at 20 °C for 60 min, ligase inactivation at 65 °C for 10 min, and nuclease treatment at 37 °C for 1 h. We purified and concentrated the library with 0.45× Ampure PB beads for size selection to collect fragments greater than 7 to 9 kb using the BluePippin/PippinHT system (Sage Science, MA; Cat #BLF7510/HPE7510). The HiFi library averaged 15 to 20 kb. It was sequenced at UC Davis DNA Technologies Core (Davis, CA) using two 8M SMRT cells, Sequel II sequencing chemistry 2.0, and 30-h movies each on a PacBio Sequel II sequencer.

We used the Omni-C™ Kit (Dovetail Genomics, CA) for Omni-C proximity sequencing according to the manufacturer's protocol with slight modifications. First, we ground muscle tissue (Sample YEWA_CCGP3; LACM Bird #122168, KLG4550, LAF9440) with a mortar and pestle while cooled with liquid nitrogen. Subsequently, chromatin was fixed in place in the nucleus. We passed the suspended chromatin solution through 100 µm and 40 µm cell strainers to remove large debris. We digested fixed chromatin under various conditions of DNase I until a suitable fragment length distribution of DNA molecules was obtained. We repaired chromatin ends, ligated a biotinylated bridge adapter, and performed proximity ligation of adapter-containing ends. After proximity ligation, crosslinks were reversed, and the DNA was purified from proteins. We treated purified DNA to remove biotin that was not internal to ligated fragments. We generated a next-generation sequencing library using an NEB Ultra II DNA Library Prep kit (New England Biolabs, MA) with an Illumina-compatible y-adapter. Then, we captured biotin-containing fragments using streptavidin beads. We split the post-capture product into two replicates before PCR enrichment to preserve library complexity, with each replicate receiving unique dual indices. The library was sequenced at the Vincent J. Coates Genomics Sequencing Lab (Berkeley, CA) on an Illumina NovaSeq 6000 PE150 (Illumina, CA). Based on a genome size of 1.26 Gb (Bay et al. 2018), we targeted 126 million base pair reads (100 million read pairs per Gb genome size).

Nuclear genome assembly

We assembled the Yellow Warbler genome 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). Then, we obtained the initial phased diploid assembly using HiFiasm (Cheng et al. 2022) with the filtered PacBio HiFi reads and the Omni-C dataset. 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 multi-resolution Omni-C matrix with a cooler (Abdennur and 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 and then we checked the contact maps for major misassemblies. In detail, if we identified a strong off-diagonal signal and a lack of signal in the consecutive genomic region in the proximity of a join made by the scaffolder, we dissolved it by breaking the scaffolds at the coordinates of the join. After this process, no further manual joins were made. Some 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). Given the similar contiguity metrics and fragmentation of both assemblies, we decided to tag the assemblies as primary and alternate, where primary is the one that overall is more complete, has better BUSCO (Benchmarking Universal Single-Copy Orthologs) scores and better k-mer completeness.

Table 1. Assembly pipeline and software used for assembly of the Yellow Warbler genome.

Purpose	Software[a]	Version
Assembly		
Adapters	HiFiAdapterFilt	Commit 64d1c7b
K-mer counting	Meryl (k = 21)	1
Estimation of genome size and heterozygosity	GenomeScope	2
De novo *assembly (contiging)*	HiFiasm (Hi-C Mode, –primary, output p_ctg.hap1, p_ctg.hap2)	0.16.1-r375
Scaffolding		
Omni-C 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+ (-db nt, -outfmt '6 qseqid staxids bitscore std' , -max_target_seqs 1, -max_hsps 1, -evalue 1e-25)	2.1
General contamination screening	BlobToolKit (PacBIo HiFi Coverage, NCBI Taxa ID = 123631, BUSCODB = aves)	2.3.3

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

Genome assembly 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 overall repeat content. To obtain general contiguity metrics, we ran QUAST (Gurevich et al. 2013). To evaluate genome quality and functional completeness, we used BUSCO (Manni et al. 2021) with the Aves ortholog database (aves_odb10) containing 8,338 genes. Base level accuracy (QV) and k-mer completeness were assessed 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 $x.y.P.Q.C$, where x = log10[contig NG50]; y = log10[scaffold NG50]; P = log10 [phased block NG50]; Q = Phred base accuracy QV (quality value); C = % genome represented by the first "n" scaffolds, following a known karyotype of $2n$ = 80 for Yellow Warbler [Bird Chromosome Database, Chromosome number data V3.0/2022—(Hobart 1991; Degrandi et al. 2020)]. Quality metrics for the notation were calculated on the primary assembly (bSetPet1.0.p).

Mitochondrial genome assembly

We assembled the mitochondrial genome of Yellow Warbler from the PacBio HiFi reads using the reference-guided pipeline MitoHiFi (Allio et al. 2020; Uliano-Silva et al. 2021). We used the mitochondrial sequence of Kirtland's Warbler (NCBI:NC_051027.1) 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.

Results

Sequencing data

The Omni-C and PacBio HiFi sequencing libraries generated 85.3 million read pairs and 2.7 million reads, respectively.

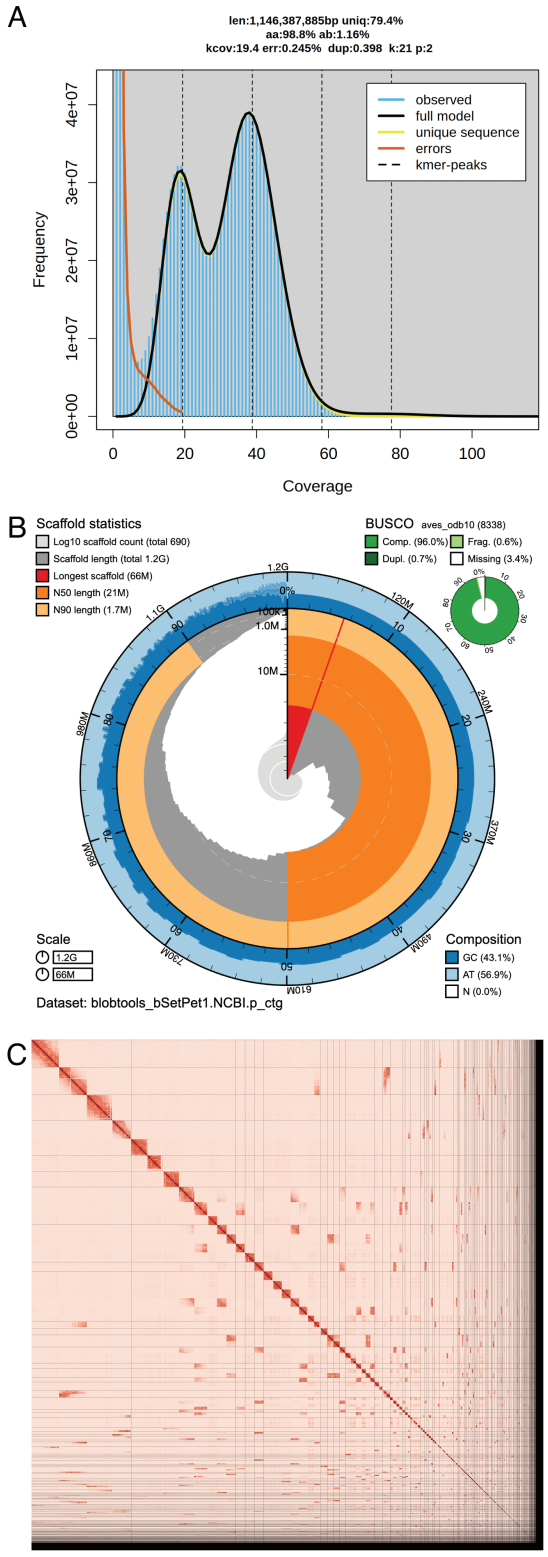

Fig. 2. Visual overview of genome assembly metrics. A) K-mer spectra output generated from PacBio HiFi data without adapters using GenomScope2.0. The bimodal pattern observed corresponds to a diploid genome. K-mers covered at lower coverage and lower frequency correspond to differences between haplotypes, 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 *Setophaga petechia* primary assembly (bSetPet1.0.p). The plot circle represents the full size of the assembly. From the inside-out, the central plot covers scaffold and length-related metrics. The central light gray spiral shows the cumulative scaffold count with a white line at each order of magnitude. 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 outer light and dark blue ring show the mean, maximum, and minimum GC vs. AT content at 0.1% intervals (Challis et al. 2020). C) 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 such regions. Scaffolds are separated by black lines and higher density corresponds to higher levels of fragmentation.

The latter yielded 40.87-fold coverage (N50 read length 17,523 bp; minimum read length 41 bp; mean read length 17,110 bp; maximum read length of 54,497 bp). Based on PacBio HiFi reads, we estimated a genome assembly size of 1.14 Gb, 79.39% sequence uniqueness (20.61% repeat content), 0.245% sequencing error rate, and 1.16% nucleotide heterozygosity rate using Genomescope2.0. The k-mer spectrum based on PacBio HiFi reads shows (Fig. 2A) a bimodal distribution with two major peaks at 19- and 39-fold coverage, where peaks correspond to heterozygous and homozygous states of a diploid species.

Nuclear genome assembly

The final assembly consists of two haplotypes tagged as primary and alternate (bSetPet1.0.p and bSetPet1.0.a). Both genome assembly sizes are similar but not equal to the estimated value from Genomescope2.0 (Fig. 2A). The primary assembly (bSetPet1.0.p) consists of 687 scaffolds spanning 1.22 Gb with contig N50 of 6.8 Mb, scaffold N50 of 21.18 Mb, longest contig of 53.52 Mb, and largest scaffold of 66.28 Mb. The alternate assembly (bSetPet1.0.a) consists of 530 scaffolds, spanning 1.24 Gb with contig N50 of 8.3Mb, scaffold N50 of 21.18 Mb, largest contig 40.02 Mb and largest scaffold of 74.56 Mb. The Omni-C contact maps suggest highly contiguous primary and alternate assemblies (Fig. 2C and Supplementary Fig. S1B). The primary assembly has a BUSCO completeness score of 96.0% using the Aves gene set, a per-base quality (QV) of 62.34, a k-mer completeness of 84.95, and a frameshift indel QV of 41.54. In comparison, the alternate assembly has a BUSCO completeness score of 93.5% using the same gene set, a per-base quality (QV) of 62.79, a k-mer completeness of 81.57, and a frameshift indel QV of 40.43.

During manual curation, we identified 13 misassemblies requiring breaking nine joins on the primary assembly and four on the alternate assembly. We were able to close a total of five gaps, three on the primary and two on the alternate assembly. We removed two contigs, one per assembly, corresponding to mitochondrial contaminants. Detailed assembly statistics are reported in Table 2, and a graphical representation of the primary assembly in Fig. 2B (see Supplementary Fig. S1A for the alternate assembly). We have deposited both assemblies on NCBI (see Table 2 and Data Availability for details).

Mitochondrial genome assembly

We assembled a mitochondrial genome with MitoHiFi. The final mitochondrial assembly has a size of 16,809 bp. The base composition of the final assembly version is *A* = 30.19%,

Table 2. Sequencing and assembly statistics and accession information for the primary and alternate assemblies of the Yellow Warbler (*Setophaga petechia*) genome.

Bio Projects and Vouchers	CCGP NCBI BioProject					PRJNA720569
	Genera NCBI BioProject					PRJNA765861
	Species NCBI BioProject					PRJNA777222
	NCBI BioSample					SAMN29044059, SAMN29044060
	Specimen identification					LACM:Birds122168
	NCBI Genome accessions		Primary			Alternate
	Assembly accession		JANCRA000000000			JANCRB000000000
	Genome sequences		GCA_024362935.1			GCA_024372515.1
Genome sequence	PacBio HiFi reads	Run	1 PACBIO_SMRT (Sequel II) run: 2.7M spots, 46.9G bases, 35.6Gb downloads			
		Accession	SRX16742538			
	Omni-C Illumina reads	Run	1 ILLUMINA (Illumina NovaSeq 6000) run: 85.3M spots, 25.8G bases, 8.6Gb			
		Accession	SRX16742539, SRX16742540			
Genome Assembly Quality Metrics	Assembly identifier (Quality code[a])					bSetPet1(6.7.P6.Q62.C)
	HiFi Read coverage[b]					40.87X
					Primary	Alternate
	Number of contigs				971	776
	Contig N50 (bp)				68,07,045	83,68,636
	Contig NG50[b]				72,19,428	89,24,963
	Longest Contigs				5,35,26,829	4,00,27,624
	Number of scaffolds				687	530
	Scaffold N50				2,11,88,473	2,11,88,473
	Scaffold NG50[b]				2,17,69,140	2,04,09,353
	Largest scaffold				6,62,88,485	7,45,62,066
	Size of final assembly				1,22,23,85,128	1,24,97,65,916
	Phased block NG50[b]				73,91,252	93,25,426
	Gaps per Gbp (# Gaps)				232(284)	197(246)
	Indel QV (Frame shift)				41.54557	40.4344848
	Base pair QV				62.3497	62.7988
						Full assembly = 62.5709
	k-mer completeness				84.9555	81.57
						Full assembly = 99.2811
	BUSCO completeness (aves) *n* =		C[c]	S[c]	D[c]	F[c]	M[c]
		P[d]	96.00%	95.30%	0.70%	0.60%	3.40%
		A[d]	93.50%	92.50%	1.00%	0.60%	5.90%
	Organelles		1 complete mitochondrial sequence			CM044545.1	

[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 *S. petechia* of 2*n* = 80 (Bird Chromosome Database, Chromosome number data V3.0/2022; Hobart 1991; Degrandi et al. 2020). Quality code for all the assembly denoted by primary assembly (bSetPet1.0.p).
[b]Read coverage and NGx statistics have been calculated based on the estimated genome size of 1.14 Gb.
[c]BUSCO Scores. Complete BUSCOs (C). Complete and single-copy BUSCOs (S). Complete and duplicated BUSCOs (D). Fragmented BUSCOs (F). Missing BUSCOs (M).
[d](P)rimary and (A)lternate assembly values.

C = 31.77%, *G* = 14.19%, *T* = 23.85%, and consists of 22 unique transfer RNAs and 13 protein-coding genes.

Discussion

Here, we present a highly contiguous genome assembly for the Yellow Warbler with two pseudo haplotypes. Our genome assemblies meet thresholds for proposed quality standards for vertebrate and avian genomes (Jarvis 2016; Kapusta and Suh 2017; Rhie et al. 2021). Compared to the existing *Setophaga* genomes, the primary Yellow Warbler genome assembly presented here has the highest BUSCO completeness (96.0% of Aves orthologs present) and the highest contig N50 (6.8 Mb). Although the Yellow-rumped and Kirtland's

Warbler genome assemblies have higher scaffold N50 values, our Yellow Warbler genome assembly has the fewest gaps greater than 5 N's (284 compared to 49K to 67K in other *Setophaga* genome assemblies), which highlights the improvement gained when using long-read sequencing technology in combination with short reads for more contiguous and complete genomes.

The reference genome presented here provides an essential resource for evolutionary research and conservation efforts in California and beyond. Future range-wide genomic analyses will facilitate investigations into the history of gene flow and divergence between the various subspecies groups in this complex (Browning 1994; Chaves et al. 2012; Machkour-M'Rabet et al. 2023). This system-wide genomic context lends itself to investigations into the genetic basis underlying both phenotypic diversity and the evolution of migration (Toews et al. 2016; Franchini et al. 2017; Delmore et al. 2020; Aguillon et al. 2021; Caballero-López et al. 2022).

Future landscape genomic analyses investigating environmental associations with genomic variation could identify loci important for local adaptation in this widespread species (Bay et al. 2018; Forester et al. 2018; Chen et al. 2022). Using this framework with future climate models will allow for predictions of how Yellow Warblers may adapt to future climate change and identify both populations that are likely to persist in and vulnerable to future climate change regimes, which will guide local conservation implementation (Fitzpatrick and Keller 2015; Shaffer et al. 2022). This will be especially important for California populations experiencing population declines and dwindling breeding habitat, which could benefit from direct conservation and management efforts (Heath and Ballard 2003; Shuford et al. 2008). Overall, the Yellow Warbler genome presented here provides a key resource for investigating phenotypic and ecological evolution and conservation in this charismatic migratory bird species.

Supplementary material

Supplementary material is available at *Journal of Heredity* Journal online.

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 Maeve Secor for help with fieldwork; Tara Luckau, Dr. Courtney Miller, and Dr. Erin Toffelmier for help with coordination and sample submission.

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]. WLET was supported by the University of California, Los Angeles, Department of Ecology and Evolutionary Biology, Lida Scott Brown Fellowship; and the National Science Foundation, Graduate Research Fellowship [DGE-2034835]. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Data availability

Data generated for this study are available under NCBI BioProject PRJNA777222. Raw sequencing data for individual with voucher LACM:122168 (NCBI BioSamples SAMN29044059, SAMN29044060) are deposited in the NCBI Short Read Archive (SRA) under SRX16742538 for PacBio HiFi sequencing data, and SRX16742539 and SRX16742540 for the Omni-C Illumina sequencing data. GenBank accessions for both primary and alternate assemblies are GCA_024362935.1 and GCA_024372515.1; and for genome sequences JANCRA000000000 and JANCRB000000000. The GenBank organelle genome assembly for the mitochondrial genome is CM044545.1. Assembly scripts and other data for the analyses presented can be found at the following GitHub repository: www.github.com/ccgproject/ccgp_assembly.

References

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

Aguillon SM, Walsh J, Lovette IJ. Extensive hybridization reveals multiple coloration genes underlying a complex plumage phenotype. Proc Biol Sci. 2021:288:20201805. doi:10.1098/rspb.2020.1805

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

Bay RA, Harrigan RJ, Underwood VL, Gibbs HL, Smith TB, Ruegg K. Genomic signals of selection predict climate-driven population declines in a migratory bird. Science. 2018:359:83–86. doi:10.1126/science.aan4380

Browning MR. A taxonomic review of *Dendroica petechia* (Yellow Warbler) (Aves: Parulinae). Proc Biol Soc Wash. 1994:107:27–51.

Caballero-López V, Lundberg M, Sokolovskis K, Bensch S. Transposable elements mark a repeat-rich region associated with migratory phenotypes of willow warblers (*Phylloscopus trochilus*). Mol Ecol. 2022:31:1128–1141. doi:10.1111/mec.16292

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. BMC Bioinf. 2009:10:421. doi:10.1186/1471-2105-10-421

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

Chavarria-Pizarro T, Gomez JP, Ungvari-Martin J, Bay R, Miyamoto MM, Kimball R. Strong phenotypic divergence in spite of low genetic structure in the endemic Mangrove Warbler subspecies (*Setophaga petechia xanthotera*) of Costa Rica. Ecol Evol. 2019:9:13902–13918. doi:10.1002/ece3.5826

Chaves JA, Parker PG, Smith TB. Origin and population history of a recent colonizer, the yellow warbler in Galápagos and Cocos Islands. J Evol Biol. 2012:25:509–521. doi:10.1111/j.1420-9101.2011.02447.x

Chen Y, Jiang Z, Fan P, Ericson PGP, Song G, Luo X, Lei F, Qu Y. The combination of genomic offset and niche modelling provides insights into climate change-driven vulnerability. Nat Commun. 2022:13:1. doi:10.1038/s41467-022-32546-z

Cheng H, Jarvis ED, Fedrigo O, Koepfli K-P, Urban L, Gemmell NJ, Li H. Haplotype-resolved assembly of diploid genomes without parental data. Nat Biotechnol. 2022:40:1332–1335. doi:10.1038/s41587-022-01261-x

Collinge SK, Holyoak M, Barr CB, Marty JT. Riparian habitat fragmentation and population persistence of the threatened valley elderberry longhorn beetle in central California. Biol Conserv. 2001:100:103–113. doi:10.1016/S0006-3207(00)00211-1

Dahl, T. E. (1990). Wetlands losses in the United States, 1780's to 1980's. Report to the Congress (PB-91-169284/XAB). St. Petersburg, FL (USA): National Wetlands Inventory. [accessed 2020 Feb 21]. https://www.osti.gov/biblio/5527872-wetlands-losses-united-states-report-congress

Davidson C, Bradley Shaffer H, Jennings MR. Declines of the California Red-Legged Frog: climate, Uv-B, habitat, and pesticides hypotheses. Ecol Appl. 2001:11:464–479. doi:10.1890/1051-0761(2001)011[0464:DOTCRL]2.0.CO;2

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. 2020:160:199–205. doi:10.1159/000507768

Delmore K, Illera JC, Pérez-Tris J, Segelbacher G, Lugo Ramos JS, Durieux G, Ishigohoka J, Liedvogel M. The evolutionary history and genomics of European blackcap migration. eLife. 2020:9:e54462. doi: 10.7554/eLife.54462

DeSaix MG, George TL, Seglund AE, Spellman GM, Zavaleta ES, Ruegg KC. Forecasting climate change response in an alpine specialist songbird reveals the importance of considering novel climate. Divers Distrib. 2022:28:2239–2254. doi:10.1111/ddi.13628

Feng S, StillerJ, 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. doi:10.1038/s41586-020-2873-9

Fink D, Auer T, Johnston A, Strimas-Mackey M, Ligocki S, Robinson O, Hochachka W, Jaromczyk L, Rodewald A, Wood C, et al. *eBird Status and Trends, Data Version: 2021*; Released: 2022. Ithaca, NY: Cornell Lab of Ornithology; 2022. doi:10.2173/ebirdst.2021

Fitzpatrick MC, Keller SR. Ecological genomics meets community-level modelling of biodiversity: mapping the genomic landscape of current and future environmental adaptation. Ecol Lett. 2015:18:1–16. doi:10.1111/ele.12376

Forester BR, Lasky JR, Wagner HH, Urban DL. Comparing methods for detecting multilocus adaptation with multivariate genotype–environment associations. Mol Ecol. 2018:27:2215–2233. doi:10.1111/mec.14584

Franchini P, Irisarri I, Fudickar A, Schmidt A, Meyer A, Wikelski M, Partecke J. Animal tracking meets migration genomics: transcriptomic analysis of a partially migratory bird species. Mol Ecol. 2017:26:3204–3216. doi:10.1111/mec.14108

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

Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, Phillippy AM, Koren S. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. PLoS Comput Biol. 2019:15:e1007273. doi:10.1371/journal.pcbi.1007273

Gibbs HL, Dawson RJG, Hobson KA. Limited differentiation in microsatellite DNA variation among northern populations of the yellow warbler: evidence for male-biased gene flow? Mol Ecol. 2000:9:2137–2147. doi:10.1046/j.1365-294X.2000.01136.x

Goloborodko A, Abdennur N, Venev S, hbbrandao, & gfudenberg. (2018). *mirnylab/pairtools: V0.2.0* [Computer software]. Zenodo. doi:10.5281/zenodo.1490831

Gopalakrishnan S, Samaniego Castruita JA, Sinding M-HS, Kuderna LFK, Räikkönen J, Petersen B, Sicheritz-Ponten T, Larson G, Orlando L, Marques-Bonet T, et al. The wolf reference genome sequence (*Canis lupus* lupus) and its implications for *Canis* spp. population genomics. BMC Genomics. 2017:18:495. doi:10.1186/s12864-017-3883-3

Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013:29:1072–1075. doi:10.1093/bioinformatics/btt086

Heath SK, Ballard G. Patterns of breeding songbird diversity and occurrence in Riparian habitats of the Eastern Sierra Nevada. In *California riparian systems: processes and floodplain management, ecology and restoration. Riparian Habitat and Floodplains Conference Proceedings, Riparian Habitat Joint Venture, Sacramento, CA*. 2003, (pp. 21-34).

Hobart HH. Comparative karyology in nine-primaried oscines (Aves); 1991. [accessed 2022 Oct 25]. https://repository.arizona.edu/handle/10150/185492

Jarvis ED. Perspectives from the Avian Phylogenomics Project: questions that can be answered with sequencing all genomes of a vertebrate class. Annu Rev Anim Biosci. 2016:4:45–59. doi:10.1146/annurev-animal-021815-111216

Kapusta A, Suh A. Evolution of bird genomes—a transposon's-eye view. Ann N Y Acad Sci. 2017:1389:164–185. doi:10.1111/nyas.13295

Kerpedjiev P, Abdennur N, Lekschas F, McCallum C, Dinkla K, Strobelt H, Luber JM, Ouellette SB, Azhir A, Kumar N, et al. HiGlass: web-based visual exploration and analysis of genome interaction maps. Genome Biol. 2018:19:125. doi:10.1186/s13059-018-1486-1

Klein NK, Brown WM. Intraspecific molecular phylogeny in the yellow warbler (*Dendroica petechia*), and implications for Avian Biogeography in the West Indies. Evolution. 1994:48:1914–1932. doi:10.2307/2410517

Korlach J, Gedman G, Kingan SB, Chin C-S, Howard JT, Audet J-N, Cantin L, Jarvis ED. De novo PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. GigaScience. 2017:6:gix085. doi:10.1093/gigascience/gix085

Krueper DJ. Effects of livestock management on Southwestern riparian ecosystems. In: Shaw DW, Finch DM, editors. Tech Coords. Desired Future Conditions for Southwestern Riparian Ecosystems: Bringing Interests and Concerns Together. 1995 Sept. 18–22, 1995; Albuquerque, NM. *General Technical Report RM-GTR-272*. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Forest and Range Experiment Station; 1996. p. 272, 281–301.

Lamichhaney S, Han F, Berglund J, Wang C, Almén MS, Webster MT, Grant BR, Grant PR, Andersson L. A beak size locus in Darwin's finches facilitated character displacement during a drought. Science. 2016:352:470–474. doi:10.1126/science.aad8786

Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, *arXiv*. 2013:1303.3997. https://doi.org/10.48550/arXiv.1303.3997.

Machkour-M'Rabet S, Santamaría-Rivero W, Dzib-Chay A, Cristiani LT, MacKinnon-Haskins B. Multi-character approach reveals a new mangrove population of the Yellow Warbler complex, *Setophaga petechia*, on Cozumel Island, Mexico. PLoS One. 2023:18:e0287425. doi:10.1371/journal.pone.0287425

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol. 2021:38:4647–4654. doi:10.1093/molbev/msab199

Mérot C, Oomen RA, Tigano A, Wellenreuther M. A roadmap for understanding the evolutionary significance of structural genomic variation. Trends Ecol Evol. 2020:35:561–572. doi:10.1016/j.tree.2020.03.002

Milot E, Gibbs HL, Hobson KA. Phylogeography and genetic structure of northern populations of the yellow warbler (*Dendroica petechia*). Mol Ecol. 2000:9:667–681. doi:10.1046/j.1365-294x.2000.00897.x

Peona V, Blom MPK, Xu L, Burri R, Sullivan S, Bunikis I, Liachko I, Haryoko T, Jønsson 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. 2021:21:263–286. doi:10.1111/1755-0998.13252

Phillips SE, Hamilton LP, Kelly PA. Assessment of habitat conditions for the Riparian Brush Rabbit on the San Joaquin River National Wildlife Refuge, California. *Endangered Species Recovery Program*. 2005.

Poff B, Koestner KA, Neary DG, Merritt D. Threats to western United States riparian ecosystems: a bibliography (RMRS-GTR-269). U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station; 2012. doi:10.2737/RMRS-GTR-269

Prasad A, Lorenzen ED, Westbury MV. Evaluating the role of reference-genome phylogenetic distance on evolutionary inference. Mol Ecol Resour. 2022:22:45–55. doi:10.1111/1755-0998.13457

Ramírez F, Bhardwaj V, Arrigoni L, Lam KC, Grüning BA, Villaveces J, Habermann B, Akhtar A, Manke T. High-resolution TADs reveal DNA sequences underlying genome organization in flies. Nat Commun. 2018:9:1. 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. doi:10.1038/s41467-020-14998-3

Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W, Fungtammasan A, Gedman GL, et al. 2020. Towards complete and error-free genome assemblies of all vertebrate species. *Nature*. 2021;592:737–746.

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. doi:10.1038/s41586-021-03451-0

Salgado-Ortiz J, Marra PP, Sillett TS, Robertson RJ. Breeding ecology of the Mangrove Warbler (*Dendroica petechia bryanti*) and comparative life history of the yellow warbler subspecies complex. Auk. 2008:125:402–410. doi:10.1525/auk.2008.07012

Sauer JR, Hines JE, Fallon JE, Pardiek KL. The North American Breeding Bird Survey, results and analysis 1966–2012 (Version 02.19). U.S. Geological Survey Patuxent Wildlife Research Center; 2014.

Shaffer HB, Toffelmier E, Corbett-Detig RB, Escalona M, Erickson B, Fiedler P, Gold M, Harrigan RJ, Hodges S, Luckau TK, et al. Landscape genomics to enable conservation actions: the California conservation genomics project. J Hered. 2022:113:577–588. doi:10.1093/jhered/esac020

Shuford WD, Gardali T, Western Field Ornithologists, California, and Department of Fish and Game. California bird species of special concern: a ranked assessment of species, subspecies, and distinct populations of birds of immediate conservation concern in California. Camarillo, CA, USA: Western Field Ornithologists; California Department of Fish and Game; 2008. http://books.google.com/books?id=9INFAQAAIAAJ

Sim SB, Corpuz RL, Simmonds TJ, Geib SM. HiFiAdapterFilt, a memory efficient read processing pipeline, prevents occurrence of adapter sequence in PacBio HiFi reads and their negative impacts on genome assembly. BMC Genomics. 2022:23:157. doi:10.1186/s12864-022-08375-1

Toews DPL, Taylor SA, Vallender R, Brelsford A, Butcher BG, Messer PW, Lovette IJ. Plumage genes and little else distinguish the genomes of hybridizing warblers. Curr Biol. 2016:26:2313–2318. doi:10.1016/j.cub.2016.06.034

Uliano-Silva M, Nunes JGF, Krasheninnikova K, McCarthy SA. *marcelauliano/MitoHiFi: Mitohifi_v2.0* [Computer software]. Zenodo; 2021. doi:10.5281/zenodo.5205678

Wellenreuther M, Bernatchez L. Eco-evolutionary genomics of chromosomal inversions. Trends Ecol Evol. 2018:33:427–440. doi:10.1016/j.tree.2018.04.002

Wilson CM, Holberton RL. Individual risk versus immediate reproductive success: a basis for latitudinal differences in the adrenocortical response to stress in yellow warblers (*Dendroica petechia*). Auk. 2004:122:378–1249. doi:10.1093/auk/121.4.1238