Journal of Heredity, 2023, **114**, 436–443
https://doi.org/10.1093/jhered/esad025
Advance access publication 29 April 2023
Genome Resources

Genome Resources

Reference genome of the black rail, *Laterallus jamaicensis*

Laurie A. Hall[1,2,3], Ian J. Wang[1,2], Merly Escalona[4], Eric Beraut[5], Samuel Sacco[5], Ruta Sahasrabudhe[6], Oanh Nguyen[6], Erin Toffelmier[7,8], H. Bradley Shaffer[7,8] and Steven R. Beissinger[1,2].

[1]Department of Environmental Science, Policy and Management, University of California, Berkeley, CA 94720, United States,
[2]Museum of Vertebrate Zoology, University of California, Berkeley, CA 94720, United States,
[3]Current address: San Francisco Bay Estuary Field Station, Western Ecological Research Center, U.S. Geological Survey, Moffett Field, CA 94035, United States
[4]Department of Biomolecular Engineering, University of California, Santa Cruz, CA 95064, United States,
[5]Department of Ecology and Evolutionary Biology, University of Califin JHornia, Santa Cruz, CA 95064, United States,
[6]DNA Technologies and Expression Analysis Core Laboratory, Genome Center, University of California, Davis, CA 95616, United States,
[7]Department of Ecology and Evolutionary Biology, University of California, Los Angeles, CA 90095, United States,
[8]La Kretz Center for California Conservation Science, Institute of the Environment and Sustainability, University of California, Los Angeles, CA 90095, United States

Address correspondence to Laurie A. Hall at the address above, or e-mail: lahall@usgs.gov.

Corresponding Editor: Beth Shapiro

Abstract

The black rail, *Laterallus jamaicensis*, is one of the most secretive and poorly understood birds in the Americas. Two of its five subspecies breed in North America: the Eastern black rail (*L. j. jamaicensis*), found primarily in the southern and mid-Atlantic states, and the California black rail (*L. j. coturniculus*), inhabiting California and Arizona, are recognized across the highly disjunct distribution. Population declines, due primarily to wetland loss and degradation, have resulted in conservation status listings for both subspecies. To help advance understanding of the phylogeography, biology, and ecology of this elusive species, we report the first reference genome assembly for the black rail, produced as part of the California Conservation Genomics Project (CCGP). We produced a de novo genome assembly using Pacific Biosciences HiFi long reads and Hi-C chromatin-proximity sequencing technology with an estimated sequencing error rate of 0.182%. The assembly consists of 964 scaffolds spanning 1.39 Gb, with a contig N50 of 7.4 Mb, scaffold N50 of 21.4 Mb, largest contig of 44.8 Mb, and largest scaffold of 101.2 Mb. The assembly has a high BUSCO completeness score of 96.8% and represents the first genome assembly available for the genus *Laterallus*. This genome assembly can help resolve questions about the complex evolutionary history of rails, assess black rail vagility and population connectivity, estimate effective population sizes, and evaluate the potential of rails for adaptive evolution in the face of growing threats from climate change, habitat loss and fragmentation, and disease.

Key words: California Conservation Genomics Project, CCGP, conservation genetics, Gruiformes, Rallidae

Introduction

The black rail, *Laterallus jamaicensis*, is a small (~30 g) bird that occupies densely-vegetated freshwater and brackish wetland habitats. It is one of the most secretive birds in North America, and knowledge of its phylogeography, biology, and ecology is limited. Two of its five subspecies breed in North America. The Eastern subspecies, *L. j. jamaicensis*, is listed as Threatened under the US Endangered Species Act and has both migratory and non-migratory populations that are distributed south of Massachusetts along the East Coast of the United States of America; along the Gulf Coast of the United States of America and Mexico; and in isolated patches throughout the Great Plains and Eastern United States of America (Watts 2016; Eddleman et al. 2020). The California subspecies, *L. j. coturniculus*, is listed as Threatened by the California Department of Fish and Wildlife and has non-migratory populations that are distributed in the Sierra Nevada foothills and San Francisco Bay area of California and the Imperial Valley region of extreme southeastern California and adjacent western Arizona and Mexico (Richmond et al. 2008; Girard et al. 2010; Eddleman et al. 2020). Loss and degradation of wetland habitats due to changes in climate and land use are thought to be the primary threats to black rail populations, but West Nile Virus has also contributed to black rail declines (Veloz et al. 2013; Roach and Barrett 2015; Watts 2016; Beissinger et al. 2022).

Black rails are classified as Rallidae, a diverse and globally distributed family with 40 genera (Kirchman et al. 2021). Despite their propensity to colonize remote islands, rails are generally considered poor flyers, and flightlessness has evolved multiple times within the family, resulting in a complex evolutionary history that has been recently revised

Received January 12, 2023; Accepted April 27, 2023

(Kirchman 2012; Garcia-R and Mazke 2021; Kirchman et al. 2021). Novel genomic data for rail taxa, including black rails, would help resolve uncertainty among phylogeographic relationships and contribute to our understanding of the evolutionary history of Rallidae.

Across the black rail's highly disjunct distribution, direct assessment of dispersal, and migratory movements remains challenging because the elusive nature of rails makes it difficult to successfully conduct mark-release-recapture studies. Further, the densely-vegetated habitats of rails can result in entanglement of telemetred individuals and often have inadequate light levels for solar charging. Therefore, previous studies have relied on data from intensive occupancy surveys, stable isotopes, and genetic markers to indirectly infer black rail movements (Hall and Beissinger 2017; Hall et al. 2018). A better understanding of black rail vagility, population connectivity, and genetic diversity would enable resource managers to select locations for wetland protection, restoration, and enhancement efforts, helping to secure critical habitat for this species. In addition, determination of management units and effective population sizes using genetic data could inform management actions because accurate census population sizes have been virtually impossible to obtain, presenting a major impediment to recovery efforts (Richmond et al. 2008; Girard et al. 2010; Watts 2016).

Here we report the first reference genome assembly for the black rail, produced as part of the California Conservation Genomics Project (CCGP; Shaffer et al. 2022). This genome will provide a resource for future black rail genomics studies, helping to advance our understanding of the phylogeography, biology, and ecology of this elusive species, including efforts by state, and local agencies to recover the species (Fiedler et al. 2022).

Methods

Biological materials

One female California black rail (*L. j. coturniculus*; Fig. 1) was captured at Spenceville Wildlife Area in Penn Valley, CA, United States of America (39.101991N, −121.291638W) on 3 August 2020 with a mist-net following the methods of Girard et al. (2010; California Department of Fish and Wildlife Permit SC-4438 to SRB). Whole blood (~150 µl) was collected from a brachial wing vein using a 26-gauge hypodermic needle and heparinized capillary tubes. Equal aliquots of blood were stored in two microtubes with 6.16 nM Na EDTA. Blood was transported on ice to a field station where it was refrigerated at 4 °C for 24 h before one aliquot was transported on ice to the University of California Davis DNA Technologies and Expression Analysis Core Laboratory (Davis, CA, United States of America); the second aliquot was transported on ice to the University of California Santa Cruz Paleogenomics Laboratory (Santa Cruz, CA, United States of America).

High molecular weight genomic DNA isolation

High molecular weight (HMW) genomic DNA (gDNA) was isolated from whole blood preserved in EDTA. 20 µl of whole blood was added to 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. Lysis was carried out at room temperature for a few hours until the solution was homogenous. The lysate was treated with 20 µg/ml RNase A at 37 °C for 30 min and cleaned with equal volumes of phenol/chloroform using phase lock gels (Quantabio, MA, United States of America; Cat # 2302830). DNA was precipitated by adding 0.4× volume of 5 M ammonium acetate and 3× volume of ice-cold ethanol. The DNA pellet was washed twice with 70% ethanol and resuspended in an elution buffer (10 mM Tris, pH 8.0). Purity of gDNA was assessed using a NanoDrop ND-1000 spectrophotometer, and a 260/280 ratio of 1.91 and 260/230 of 2.13 were observed. DNA yield (150 µg total) was quantified using a Qubit 2.0 Fluorometer (Thermo Fisher Scientific, MA, United States of America). Integrity of the HMW gDNA was verified on a Femto pulse system (Agilent Technologies, CA, United States of America), and 71.9% of the DNA was found in fragments larger than 120 Kb in length.

Fig. 1. Photo of a California black rail (*Laterallus jamaicensis coturniculus*). Credit: Orien Richmond.

HiFi library preparation and sequencing

The HiFi SMRTbell library was constructed using the SMRTbell Express Template Prep Kit v2.0 (Pacific Biosciences—PacBio, CA, United States of America; 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. In detail, each shearing was added to a hydro tube (Diagenode, Denville, NJ, Cat. No. C30010018) for attachment to a long hydropore (Diagenode, Cat. No. E07010002) for shearing at speeds 34–35 with specific concentrations and volumes required by the Megaruptor software (Diagenode, Cat# B06010003). Sizing of each input shearing was verified by Femto Pulse (Agilent) before pooling and concentrating for next step (Supplementary Table 1). The sheared gDNA was concentrated using 0.45× of AMPure PB beads (PacBio, CA, United States of America; Cat. #100-265-900) for the removal of single-strand overhangs at 37 °C for 15 min, followed by further enzymatic steps of DNA damage repair at 37 °C for 30 min, end repair, and A-tailing at 20 °C for 10 min and 65 °C for 30 min, ligation of overhang adapter v3 at 20 °C for 60 min, and 65 °C for 10 min to inactivate the ligase, then nuclease treated at 37 °C for 1 h. The SMRTbell library was purified and concentrated with 0.45× Ampure PB beads (PacBio, CA, United States of America; Cat. #100-265-900) for size selection using the BluePippin system (Sage Science, MA, United States of America; Cat #BLF7510) to collect fragments greater than 3–5 Kb. The 15–20 kb average HiFi SMRTbell library was sequenced at the University of California Davis DNA Technologies and Expression Analysis Core Laboratory (Davis, CA, United States of America) using two 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 a Dovetail Omni-C Kit (Dovetail Genomics, CA, United States of America) according to the manufacturer's protocol with slight modifications. Briefly, chromatin was fixed in place in the nucleus. Fixed chromatin was digested with DNase I, then extracted. Chromatin ends were repaired and ligated to a biotinylated bridge adapter followed by proximity ligation of adapter-containing ends. After proximity ligation, crosslinks were reversed, and the DNA was purified from proteins. Purified DNA was treated to remove biotin that was not internal to ligated fragments, and an NGS library was generated using an NEB Ultra II DNA Library Prep kit (New England Biolabs—NEB, MA, United States of America) with an Illumina compatible y-adaptor. Biotin-containing fragments were then captured using streptavidin beads. The post-capture product was split into two replicates prior to PCR enrichment to preserve library complexity with each replicate receiving unique dual indices. The library was prepared at the University of California Santa Cruz Paleogenomics Laboratory (Santa Cruz, CA, United States of America) and sequenced at the Vincent J. Coates Genomics Sequencing Laboratory at University of California Berkeley (Berkeley, CA, United States of America) on an Illumina NovaSeq platform to generate approximately 185 million 2 × 150 bp read pairs.

Nuclear genome assembly

We assembled the California black rail genome following the CCGP assembly protocol Version 2.0, outlined in Table 1, which uses PacBio HiFi reads and Omni-C data for the generation of high quality and highly contiguous nuclear genome assemblies while minimizing manual curation. First, we removed remnant adapter sequences from the PacBio HiFi dataset using HiFiAdapterFilt (Sim et al. 2022) and from the resulting HiFi dataset we generated the initial diploid assembly using HiFiasm (Cheng et al. 2022). The diploid assembly consists of two pseudo-haplotypes (primary and alternate), where the primary assembly is more complete and consists of longer phased blocks, and the alternate consists of haplotigs (contigs in the same haplotype) in heterozygous regions, is not as complete, and is more fragmented. Given these characteristics, it cannot be considered on its own but as a complement of the primary assembly (https://lh3.github.io/2021/04/17/concepts-in-phased-assemblies, https://www.ncbi.nlm.nih.gov/grc/help/definitions/).

Next, 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 scaffolded both assemblies using the Omni-C data with SALSA (Ghurye et al. 2017, 2018).

We generated Omni-C contact maps for both assemblies by aligning the Omni-C data against the corresponding assembly with BWA-MEM (Li 2013), identified ligation junctions, and generated Omni-C pairs using pairtools (Goloborodko et al. 2018). We generated a multi-resolution 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. If in the proximity of a join that was made by the scaffolder we identified a strong signal off-diagonal and lack of signal in the consecutive genomic region, we marked this join. All marked joins were "dissolved", meaning that we broke the scaffolds at the coordinates of these joins. After this, no further joins were made. Using the PacBio HiFi reads and YAGCloser (https://github.com/merlyescalona/yagcloser), we closed some of the remaining gaps generated during scaffolding. We then checked for contamination using the BlobToolKit Framework (Challis et al. 2020). Finally, we trimmed remnants of sequence adaptors and mitochondrial contamination identified during the NCBI contamination screening upon submission of the genome assembly to GenBank.

Mitochondrial genome assembly

We assembled the mitochondrial genome of the California black rail from the PacBio HiFi reads using the reference-guided pipeline MitoHiFi (https://github.com/marcelauliano/MitoHiFi) (Allio et al. 2020). The mitochondrial sequence of *Laterallus spilonota* (NCBI:NC_056095.1) was used as the starting reference sequence. After completion of the nuclear genome, we searched for matches of the resulting mitochondrial assembly sequence in the nuclear genome assembly using BLAST+ (Camacho et al. 2009) and filtered out contigs and scaffolds from the nuclear genome with a percentage of sequence identity >99% and size smaller than the mitochondrial assembly sequence.

Genome size estimation and quality assessment

We generated k-mer counts from the PacBio HiFi reads using meryl (https://github.com/marbl/meryl). The k-mer database was then used in GenomeScope 2.0 (Ranallo-Benavidez et al. 2020) to estimate genome features including genome size, heterozygosity, and repeat content. To obtain general contiguity metrics, we ran QUAST (Gurevich et al. 2013). To evaluate genome quality and completeness we used BUSCO (Manni et al. 2021) with the Aves ortholog database (aves_odb10), which contains 8,338 genes. Assessment of base level accuracy (QV) and k-mer completeness were performed using the previously generated meryl database and merqury (Rhie et al. 2020). We further estimated genome assembly accuracy via BUSCO gene set frame shift analysis using the pipeline described in Korlach et al. (2017). Following data availability and quality metrics established by Rhie et al. (2021), we used the derived genome quality notation *x.y.Q.C*, where, x = log10[contig NG50]; y = log10[scaffold NG50]; Q = Phred base accuracy QV (quality value); C = % genome represented by the first "*n*"

Table 1. Reference genome assembly protocol version 2.0 used by the California Conservation Genomics Project.

Assembly	Software and options[§]	Version
Filtering PacBio HiFi adapters	HiFiAdapterFilt	Commit 64d1c7b
K-mer counting	Meryl (*k* = 21)	1
Estimation of genome size and heterozygosity	GenomeScope	2
De novo assembly (contiging)	HiFiasm (Hi-C mode, –primary, p_ctg and a_ctg output)	0.16.1-r375
Remove low-coverage, duplicated contigs	purge_dups	1.2.6
Scaffolding		
Omni-C Scaffolding	SALSA (-DNASE, -i 20, -p yes)	2
Gap closing	YAGCloser (-mins 2 -f 20 -mcc 2 -prt 0.25 -eft 0.2 -pld 0.2)	Commit 20e2769
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 PretextMap PretextView PretextSnapshot	2.1.11 0.1.4 0.1.5 0.03
Organelle assembly		
Mitogenome assembly	MitoHiFi (-r, -p 50, -o 1	Commit c06ed3e
Genome quality assessment		
Basic assembly metrics	QUAST (--est-ref-size)	5.0.2
Assembly completeness	BUSCO (-m geno, -l actinopterygii)	5.0.0
	Merqury	2022-01-29
Contamination screening		
General contamination screening	BlobToolKit	2.3.3
Local sequence alignment	BLAST+	2.1

[§] Options detailed for non-default parameters.

scaffolds, following a known karyotype of 2*n* = 76 inferred from another species in the same genus, *Laterallus viridis* (Bird Chromosome Database V3.0/2022). Quality metrics for the notation were calculated on the primary assembly.

Results

The Omni-C and PacBio HiFi sequencing libraries generated 129.8 million read pairs and 3.1 million reads, respectively. The latter yielded 39.4-fold coverage (N50 read length 15,038 bp; minimum read length 43 bp; mean read length 14,930 bp; maximum read length of 51,509 bp) based on the Genomescope 2.0 genome size estimation of 1.19 Gb. Using Genomescope 2.0, we estimated 0.182% sequencing error rate and 0.856% nucleotide heterozygosity rate from the k-mer spectrum based on PacBio HiFi reads. The k-mer spectrum shows a bimodal distribution with two major peaks at ~19- and 38-fold coverage, where peaks correspond to homozygous and heterozygous states of a diploid species (Fig. 2A). The distribution presented in this k-mer spectrum was similar to that of the *C. californica*, which had a heterozygosity rate (0.73%) slightly lower than what we observed for the black rail (Benham et al. 2023).

The final assembly (bLatJam1) consists of two pseudohaplotypes, primary, and alternate. Both genome sizes were similar to the estimated value from Genomescope 2.0 (Fig. 2A). The primary assembly is more contiguous and consists of 964 scaffolds spanning 1.39 Gb with a contig N50 of 7.4 Mb, scaffold N50 of 21.4 Mb, largest contig of 44.8 Mb, and largest scaffold of 101.2 Mb. In contrast, the alternate assembly consists of 5,235 scaffolds, spanning 1.21 Mb with a contig N50 of 0.78 Mb, scaffold N50 of 5.8 Mb, largest contig of 6.6 Mb, and largest scaffold of 56.7 Mb. Detailed assembly metrics are reported in Table 2 and Fig. 2B. The curation process for the primary assembly required us to break 22 of the joins generated during scaffolding corresponding to misassemblies. We closed three gaps on the primary assembly and 12 on the alternate. Based on NCBI feedback upon submission, we removed 2 contigs from the alternate assembly that were exact duplicated sequences and trimmed a single 73 bp long remainder sequencing adapter. Contact maps for

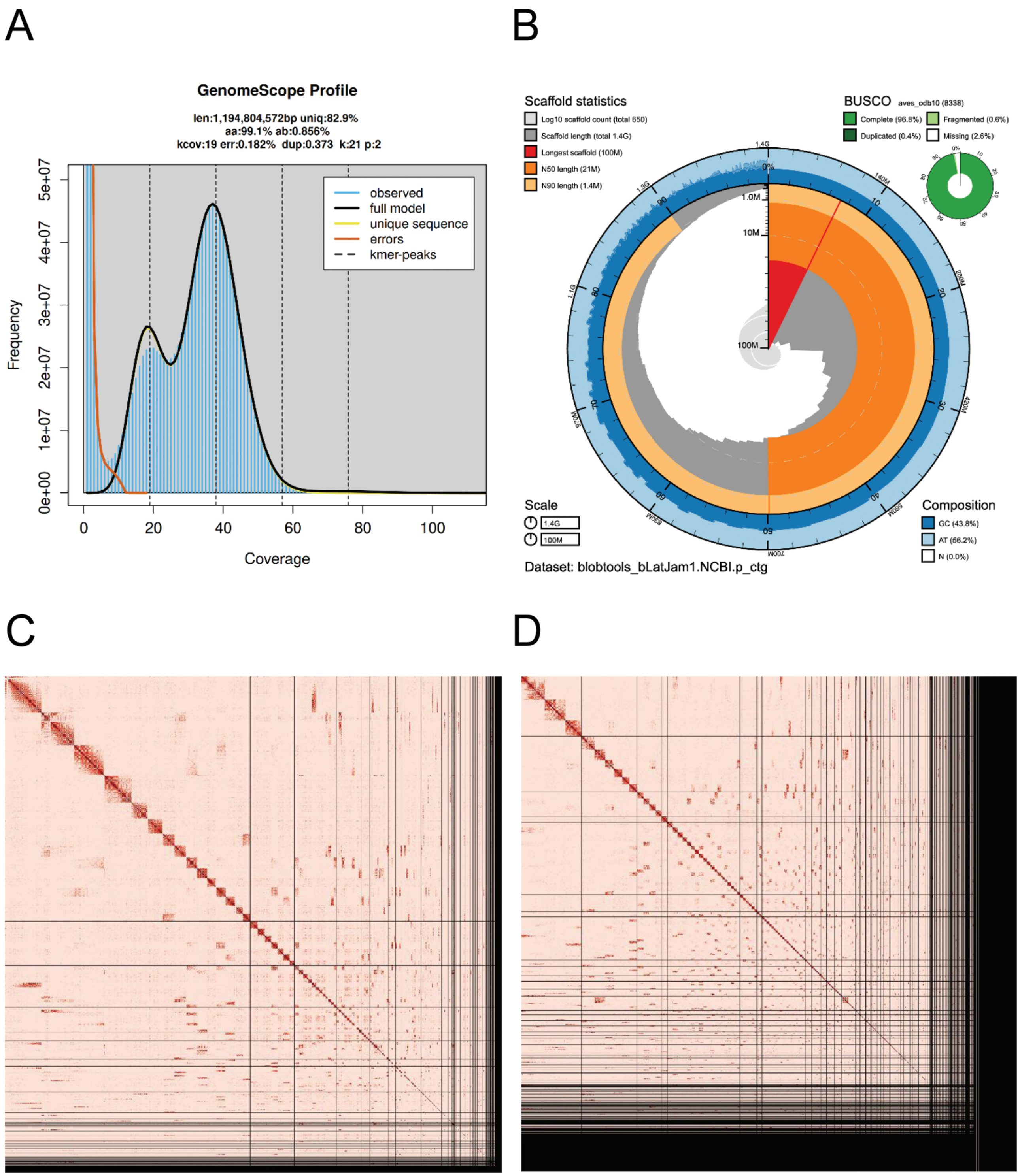

Fig. 2. Visual overview of genome assembly metrics. (A) K-mer spectrum generated from PacBio HiFi data without adapters using GenomeScope2.0. The bimodal pattern observed corresponds to a diploid genome. K-mers covered at lower coverage and lower frequency correspond to differences between haplotypes, whereas the higher coverage and higher frequency k-mers correspond to the similarities between haplotypes. (B) BlobToolKit Snail plot showing a graphical representation of the quality metrics presented in Table 2 for the black rail primary assembly (bLatJam1.0.p). 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. Omni-C contact maps for the primary (C) and alternate (D) genome assembly generated with PretextSnapshot. Omni-C contact maps translate proximity of genomic regions in 3D space to contiguous linear organization. Each cell in the contact map corresponds to sequencing data supporting the linkage (or join) between two of such regions. Black lines differentiate scaffolds.

Table 2. Reference genome assembly metrics for the black rail, *Laterallus jamaicensis*, genome assembled by the California Conservation Genomics Project.

Bio Projects and vouchers	**CCGP NCBI BioProject**			**PRJNA720569**			
	Genera NCBI BioProject			PRJNA765848			
	Species NCBI BioProject			PRJNA777185			
	NCBI BioSample			SAMN24505262			
	Specimen identification			168171201			
	NCBI Genome accessions			**Primary**		**Alternate**	
	Assembly accession			JAKCOX000000000		JAKCOY000000000	
	Genome sequences			GCA_022605575.1		GCA_022605925.1	
Genome sequence	PacBio HiFi reads	Run		1 PACBIO_SMRT (Sequel II) run: 3.2M spots, 47.1G bases, 21Gb			
		Accession		SRX14572910			
	Omni-C Illumina reads	Run		1 ILLUMINA (Illumina NovaSeq 6000) run: 129.9M spots, 39.2G bases, 12.7Gb			
		Accession		SRX14572911, SRX14572912			
Genome Assembly Quality Metrics	Assembly identifier (Quality code[*])			bLatJam1(6.7.Q61.C68)			
	HiFi Read coverage[**]			37.69X			
				Primary		**Alternate**	
	Number of contigs			964		5,237	
	Contig N50 (bp)			7,445,194		781,495	
	Contig NG50 (bp)[**]			9,465,376		797,288	
	Longest Contigs			44,858,681		6,618,197	
	Number of scaffolds			645		3,565	
	Scaffold N50 (bp)			21,403,927		5,828,180	
	Scaffold NG50 (bp)[**]			25,730,285		5,840,391	
	Largest scaffold			101,238,236		56,777,143	
	Size of final assembly (bp)			1,391,405,863		1,211,411,847	
	Gaps per Gbp (#Gaps)			222 (319)		1,380 (1,672)	
	Indel QV (Frame shift)			40.86		41.14	
	Base pair QV			61.5739		62.1354	
						Full assembly = 61.8262	
	k-mer completeness			92.6538		78.182	
				Full assembly = 99.5403			
	BUSCO completeness (Aves) *N* = 8338		C	S	D	F	M
		P[***]	96.80%	96.40%	0.40%	0.60%	2.60%
		A[***]	85.10%	84.40%	0.70%	1.00%	13.90%
	Organelles			1 Complete mitochondrial sequence CM040151.1			

[*]Assembly quality code *x*.*y*.*Q*.C derived notation, from (Rhie et al. 2021). *x* = log10[contig NG50]; *y* = log10[scaffold NG50]; *Q* = Phred base accuracy QV (Quality value); C = % genome represented by the first "*n*" scaffolds, following a known karyotype of 2*n* = 76 inferred from other species in the same genus, *Laterallus viridis* (Bird Chromosome Database V3.0/2022). BUSCO scores. (C)omplete and (S)ingle; (C)omplete and (D)uplicated; (F)ragmented and (M)issing BUSCO genes. *n*, number of BUSCO genes in the set/database. Bp: base pairs.
[**]Read coverage and NGx statistics have been calculated based on the estimated genome size of 1.19 Gb.
[***]P(rimary) and (A)lternate assembly values.

both assemblies show some level of fragmentation, where the primary assembly was clearly more contiguous than the alternate assembly, but also little evidence of inversions and translocations within contigs (Fig. 2C and D). The primary assembly has a BUSCO completeness score of 96.8% using the Aves gene set, a per base quality (QV) of 61.57, a k-mer completeness of 92.65, and a frameshift indel QV of 40.85. The alternate assembly has a BUSCO completeness score of 85.1% using the Aves gene set, a per base quality (QV) of 62.13, a k-mer completeness of 78.18, and a frameshift indel QV of 41.14. We have deposited scaffolds corresponding to both primary and alternate assemblies on NCBI (see Table 2 and Data availability for details).

The final mitochondrial assembly generated with MitoHiFi is 17,042 bp in length, with a final base composition *A* = 33.71%, *C* = 29.02%, *G* = 13.21%, and *T* = 24.06%. The final assembly consists of 22 unique transfer RNAs and 13 protein coding genes.

Discussion

The genome assembly presented here for the black rail is one of seven publicly available genome assemblies in the family Rallidae and the only genome assembly available for the genus *Laterallus*. The 1.39 Gb black rail genome was similar in size to those reported for other species of Rallidae (mean = 1.2 Gb, range = 1.1–1.4 Gb), with above average contig N50 of 7.4 Mb (Rallidae mean = 4.7 Mb, range = 0.01–13.5 Mb), and scaffold N50 of 21.4 (Rallidae mean = 20.8 Mb, range = 0.04–71.6 Mb). The GC composition of the black rail genome (44%) was also similar to the average GC composition for Rallidae (mean = 43%, range = 43–44%).

This black rail genome assembly is a critical resource to improve our understanding of rail phylogeography, biology, and ecology. Comparisons of genomic data among rail taxa have helped resolve phylogeographic relationships, and this black rail genome will advance future efforts (Kirchman 2012; Garcia-R et al. 2020; Garcia-R and Mazke 2021; Kirchman et al. 2021). For example, a recent study by Stervander et al. (2019) suggested that the Inaccessible Island rail (*Atlantisia rogersi*) was phylogenetically contained within the genus *Laterallus* and was closely related to the black rail. In addition, the black rail mitochondrial genome presented here was mapped to the genome of the Galapagos crake (*L. spilonota)*, the nearest sister taxon of the black rail, with a mean crown group age of 1.3 million years (Chaves et al. 2020).

The black rail has a highly disjunct distribution in North America with both migratory and non-migratory populations (Richmond et al. 2008; Girard et al. 2010; Watts 2016; Eddleman et al. 2020). Conservation genetic studies of species with isolated populations such as the black rail are particularly valuable for improving our understanding of how habitat loss and fragmentation affect genetic diversity, local adaptation, and vulnerability to environmental change. This genome assembly provides a key resource for population genetic studies of black rail vagility, population connectivity, and genetic diversity. Future studies will further aid conservation efforts by informing molecular model estimates of black rail movements, effective population sizes, and genes involved in adaptive evolution in the face of growing threats, including climate change, wetland loss, and disease.

Supplementary Material

Supplementary material can be found at http://www.jhered.oxfordjournals.org/.

Acknowledgments

We thank E. Burkett, J. Garcia, and C. Hirano for assistance with permits and S. Peterson for assistance in the field. PacBio Sequel II library prep and sequencing was carried out at the DNA Technologies and Expression Analysis Core Laboratories at the University of California 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 DNA Technologies and Expression Analysis Core Laboratories and the University of California Santa Cruz (UCSC) Paleogenomics Laboratory for their diligence and dedication to generating high quality sequence data. Two anonymous reviewers provided comments to improve the manuscript. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the US Government.

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].

Data availability

Data generated for this study are available under NCBI BioProject PRJNA777185. Raw sequencing data for this sample (NCBI BioSample SAMN24505262) are deposited in the NCBI Short Read Archive (SRA) under SRX14572910 for PacBio HiFi data and SRX14572911 and SRX14572912 for Omni-C Illumnia short-read data. 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(1):311–316. doi:10.1093/bioinformatics/btz540

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. doi:10.1111/1755-0998.13160

Beissinger SR, Peterson SM, Hall LA, Van Schmidt ND, Tecklin J, Risk BB, Kilpatrick AM. Stability of patch-turnover relationships under equilibrium and non-equilibrium metapopulation dynamics driven by biogeography. *Ecol Lett* 2022:25:2372–2383.

Benham PM, Cicero C, Escalona M, Beraut E, Marimuthu MPA, Nguyen O, Nachman MW, Bowie RCK. A highly contiguous genome assembly for the California quail (*Callipepla californica*). *J Hered*. 2023:esad008. doi:10.1093/jhered/esad008

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. *BMC Bioinf*. 2009:10:421

Challis R, Richards E, Rajan J, Cochrane G, Blaxter M. BlobToolKit—interactive quality assessment of genome assemblies. *G3 Genes Genomes Genet*. 2020:10(4):1361–1374. doi:10.1534/g3.119.400908

Chaves JA, Martinez-Torres PJ, Depino EA, Espinoza-Ulloa S, García-Loor J, Beichman AC, Stervander M. Evolutionary history of the galápagos rail revealed by ancient mitogenomes and modern samples. *Diversity*. 2020:12(11):4251–4215. doi:10.3390/d12110425

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

Eddleman WR, Flores RE, Legare M. Black Rail (*Laterallus jamaicensis*), version 1.0. In: Poole AF, Gill FB, editors. *Birds of the World*. Ithaca, NY, USA: Cornell Lab of Ornithology; 2020

Fiedler PL, Erickson B, Esgro M, Gold M, Hull JM, Norris J, Shaffer HB. Seizing the moment: the opportunity and relevance of the

California Conservation Genomics Project to state and federal conservation policy. *J Hered*. 2022:113(6):589–596. doi:10.1093/jhered/esac046

Garcia-R JC, Lemmon EM, Lemmon AR, French N. Phylogenomic reconstruction sheds light on new relationships and timescale of rails (Aves: Rallidae) evolution. *Diversity*. 2020:12(2):70. doi:10.3390/d12020070

Garcia-R JC, Matzke NJ. Trait-dependent dispersal in rails (Aves: Rallidae): historical biogeography of a cosmopolitan bird clade. *Mol Phylogenet Evol*. 2021:159(November 2020):107106. doi:10.1016/j.ympev.2021.107106

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. doi:10.1186/s12864-017-3879-z

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 2019:15(8):e1007273. doi:10.1371/journal.pcbi.1007273

Girard P, Takekawa JY, Beissinger SR. Uncloaking a cryptic, threatened rail with molecular markers: origins, connectivity and demography of a recently-discovered population. *Conserv Genet*. 2010:11(6):2409–2418. doi:10.1007/s10592-010-0126-4

Goloborodko, A., Abdennur, N., Venev, S., hbbrandao, & gfudenberg. (2018). *mirnylab/pairtools: v0.2.0*. Zenodo. doi: 10.5281/zenodo.1490831

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. doi:10.1093/bioinformatics/btaa025

Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. 2013:29(8):1072–1075. doi:10.1093/bioinformatics/btt086

Hall LA, Beissinger SR. Inferring the timing of long-distance dispersal between rail metapopulations using genetic and isotopic assignments. *Ecol Appl*. 2017:27(1):208–218. doi:10.1002/eap.1432

Hall LA, Van Schmidt ND, Beissinger SR. Validating dispersal distances inferred from autoregressive occupancy models with genetic parentage assignments. *J Anim Ecol*. 2018:87(3):691–702. doi:10.1111/1365-2656.12811

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

Kirchman JJ. Speciation of flightless rails on Islands: a DNA-based phylogeny of the typical rails of the Pacific. *Auk*. 2012:129(1):56–69. doi:10.1525/auk.2011.11096

Kirchman JJ, Rotzel McInerney N, Giarla TC, Olson SL, Slikas E, Fleischer RC. Phylogeny based on ultra-conserved elements clarifies the evolution of rails and allies (Ralloidea) and is the basis for a revised classification. *Ornithology*. 2021:138(4):ukab042. doi:10.1093/ornithology/ukab042

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(10):1–16. doi: 10.1093/gigascience/gix085

Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*: Genomics (2013). Retrieved from http://arxiv.org/abs/1303.3997, 2013, preprint: not peer reviewed

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.

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*. 2018:9:189. 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. 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, Kim J, et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature*. 2021:592:737–746. doi:10.1038/s41586-021-03451-0

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. doi:10.1186/s13059-020-02134-9

Richmond OM, Tecklin J, Beissinger SR. Distribution of California black rails in the Sierra Nevada Foothills. *J Field Ornithol*. 2008:79(4):381–390. doi:10.1111/j.1557-9263.2008.00195.x

Roach NS, Barrett K. Managed habitats increase occupancy of black rails (*Laterallus jamaicensis*) and may buffer impacts from sea level rise. *Wetlands*. 2015:35(6):1065–1076. doi:10.1007/s13157-015-0695-6

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(6):577–588. doi:10.1093/jhered/esac020

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 Genom*. 2022:23:157. doi:10.1186/s12864-022-08375-1

Stervander M, Ryan PG, Melo M, Hansson B. The origin of the world's smallest flightless bird, the Inaccessible Island Rail Atlantisia rogersi (Aves: Rallidae). *Molecular Phylogenetics and Evolution*. 2019:130:92-98. doi:10.1016/j.ympev.2018.10.007.

Veloz SD, Nur N, Salas L, Johgsomjit D, Wood J, Stralberg D, Ballard G. Modeling climate change impacts on tidal marsh birds: restoration and conservation planning in the face of uncertainty. *Ecosphere*. 2013:4(4):1–25

Watts BD. Status and distribution of the eastern black rail along the Atlantic and Gulf Coasts of North America. *The Center for Conservation Biology Technical Report Series: CCBTR-16-09*. 2016. College of William and Mary & Virginia Commonwealth University, Williamsburg, VA. 148 pp.