Title
Reference genome of the Virginia rail, Rallus limicola

Journal
Journal of Heredity, 114(4)

ISSN
0022-1503

Authors
Hall, Laurie A
Wang, Ian J
Escalona, Merly
et al.

Publication Date
2023-06-22

DOI
10.1093/jhered/esad026

Journal of Heredity, 2023, **114**, 428–435
https://doi.org/10.1093/jhered/esad026
Advance access publication 27 April 2023
Genome Resources

American Genetic Association

OXFORD

Genome Resources

Reference genome of the Virginia rail, *Rallus limicola*

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], Steven R. Beissinger[1,2]

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

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

Corresponding Editor: Beth Shapiro

Abstract

The Virginia rail, *Rallus limicola*, is a member of the family Rallidae, which also includes many other species of secretive and poorly studied wetland birds. It is recognized as a single species throughout its broad distribution in North America where it is exploited as a game bird, often with generous harvest limits, despite a lack of systematic population surveys and evidence of declines in many areas due to wetland loss and degradation. To help advance understanding of the phylogeography, biology, and ecology of this elusive species, we report the first reference genome assembly for the Virginia 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 0.191%. The assembly consists of 1,102 scaffolds spanning 1.39 Gb, with a contig N50 of 11.0 Mb, scaffold N50 of 25.3 Mb, largest contig of 45 Mb, and largest scaffold of 128.4 Mb. It has a high BUSCO completeness score of 96.9% and represents the first genome assembly available for the genus *Rallus*. This genome assembly will help resolve questions about the complex evolutionary history of rails and evaluate the potential of rails for adaptive evolution in the face of growing threats from climate change and habitat loss and fragmentation. It will also provide a valuable resource for rail conservation efforts by quantifying Virginia rail vagility, population connectivity, and effective population sizes.

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

Introduction

The Virginia rail, *Rallus limicola*, is a medium-sized (~85 g) bird that occupies densely vegetated freshwater and brackish marsh habitats. It is a member of the family Rallidae, a large family (152 species in 45 genera) which includes several species of secretive and poorly understood marsh birds. It is recognized as a single species throughout its broad distribution across North America, with migratory populations breeding across the continent from central Canada to the central United States and wintering in the southern United States and Mexico (Fournier et al. 2017b; Conway 2020). In addition, nonmigratory populations occur in some coastal and inland regions but are primarily distributed along the East and West coasts of the United States and inland populations in southern California, Nevada, Arizona, and New Mexico, United States (Fournier et al. 2017b; Conway 2020). The Virginia rail is common in North America and is not a species of conservation concern. It is considered a game bird with generous harvest limits throughout much of its distribution, despite a lack of systematic population surveys and evidence of declines in many areas due to wetland loss and degradation (Conway 2020).

Direct assessment of dispersal and migratory movements across the Virginia rail's broad geographic distribution remains challenging because the elusive nature of rails makes it difficult to mark and recapture individuals. Further, the densely vegetated habitats of rails often have inadequate light levels for solar charging and can result in entanglement of telemetry-marked individuals. Previous studies have relied on data from intensive occupancy surveys, citizen science efforts, and stable isotopes to indirectly infer dispersal and migratory movements (Fournier et al. 2017a, b; Beissinger et al. 2022). A better understanding of Virginia rail vagility,

Received January 12, 2023; Accepted April 27, 2023

population connectivity, and genetic diversity would enable resource managers to select optimal locations for wetland protection, restoration, and enhancement efforts, helping to secure the most important habitat for this declining species. In addition, determination of management units and effective population sizes using genetic data will help inform harvest limits and other management actions that fall under state and federal jurisdictions (Conway 2020; Fiedler et al. 2022).

Here, we report the first reference genome assembly for the Virginia rail, produced as part of the California Conservation Genomics Project (CCGP; Fiedler et al. 2022; Shaffer et al. 2022). This genome will provide a resource for future Virginia rail genomics studies, helping to advance our understanding of the phylogeography, biology, and ecology of this wide-ranging, but elusive and declining species.

Methods

Biological materials

One Virginia rail (Figure 1) was captured at Spenceville Wildlife Area in Penn Valley, California, United States (39.094599 N, −121.282184 W) on 4 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 2 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 1 aliquot was transported on ice to the University of California Davis DNA Technologies and Expression Analysis Core Laboratory (Davis, California), and the second aliquot was transported on ice to the University of California Santa Cruz Paleogenomics Laboratory (Santa Cruz, California).

High molecular weight genomic DNA isolation

High molecular weight (HMW) genomic DNA (gDNA) was isolated from whole blood preserved in EDTA. Twenty microliter 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, Massachusetts; 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 1.83 and 260/230 of 2.35 were observed. DNA yield (120 µg total) was quantified using a Qubit 2.0 Fluorometer (Thermo Fisher Scientific, Massachusetts). Integrity of the HMW gDNA was verified on a Femto pulse system (Agilent Technologies, California), and 95% of the DNA was found in fragments larger than 120 kb in length.

HiFi library preparation and sequencing

The HiFi SMRTbell library was constructed using the SMRTbell Express Template Prep Kit v2.0 (Pacific Biosciences—PacBio, California; Cat. #100-938-900) according to the manufacturer's instructions. HMW gDNA

Fig. 1. Photo of a Virginia rail (*Rallus limicola*). Credit: Orien Richmond.

was sheared to a target DNA size distribution between 15 to 20 kb. In detail, each shearing was added to a hydro tube (Diagenode, Denville, New Jersey, Cat. No. C30010018) for attachment to a long hydropore (Diagenode, Cat. No. E07010002) for shearing at speed 34 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, California; 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, California; Cat. #100-265-900) for size selection using the BluePippin system (Sage Science, Massachusetts; Cat #BLF7510) to collect fragments greater than 3 to 5 kb. The 15 to 20 kb average HiFi SMRTbell library was sequenced at the University of California Davis DNA Technologies and Expression Analysis Core Laboratory (Davis, California) 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, California) 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, 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 prepared at the University of California Santa Cruz Paleogenomics Laboratory (Santa Cruz, California) and sequenced at the Vincent J. Coates Genomics Sequencing Laboratory at University of California Berkeley (Berkeley, California) on an Illumina NovaSeq platform to generate approximately 130 million 2 × 150 bp read pairs.

Nuclear genome assembly

We assembled the Virginia rail genome following the CCGP assembly protocol Version 2.0, as 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 adapter-trimmed HiFi reads we generated the initial diploid assembly using HiFiasm (Cheng et al. 2021). The diploid assembly consists of 2 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, the alternate assembly cannot be considered on its own but as a complement of the primary assembly (https://lh3.github.io/2021/04/17/concepts-in-phased-assemblies, https://www.ncbi.nlm.nih.gov/grc/help/definitions/).

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, 2019).

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 Virginia 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 *Rallus aquaticus* (NCBI:NC_041578.1) 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

Table 1. Reference genome assembly protocol version 2.0 used by the CCGP.

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, 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

[a]Options detailed for nondefault parameters.

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 [Version 5.0.0] (Manni et al. 2021) with the Aves ortholog database (aves_odb10) which contains 8,338 genes. Assessments of base level accuracy (QV) and k-mer completeness were performed using the previously generated meryl database and merqury [V 2020-01-29] (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). 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" scaffolds, following a known karyotype of 2n = 76 inferred from another species in the same family, *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 184.4 million read pairs and 3.6 million reads, respectively. The latter yielded 40.5-fold coverage (N50 read length 13,738 bp; minimum read length 43 bp; mean read length 13,622 bp; maximum read length of 49,513 bp) based on the Genomescope 2.0 genome size estimation of 1.2 Gb. Using Genomescope 2.0, we estimated 0.191% sequencing error rate and 0.867% nucleotide heterozygosity rate from the k-mer spectrum based on PacBio HiFi reads. The k-mer spectrum shows a bimodal distribution with 2 major peaks at ~38- and 78-fold coverage, where peaks correspond to homozygous and heterozygous states of a diploid species (Figure 2A). The distribution presented in this k-mer spectrum was similar to that of the *Callipepla californica*, with the heterozygosity rate for the Virginia rail being slightly greater than the 0.73% rate for *C. californica* (Benham et al. 2023).

The final assembly (bRalLim1) consists of 2 pseudo-haplotypes, primary and alternate. Both genome sizes were

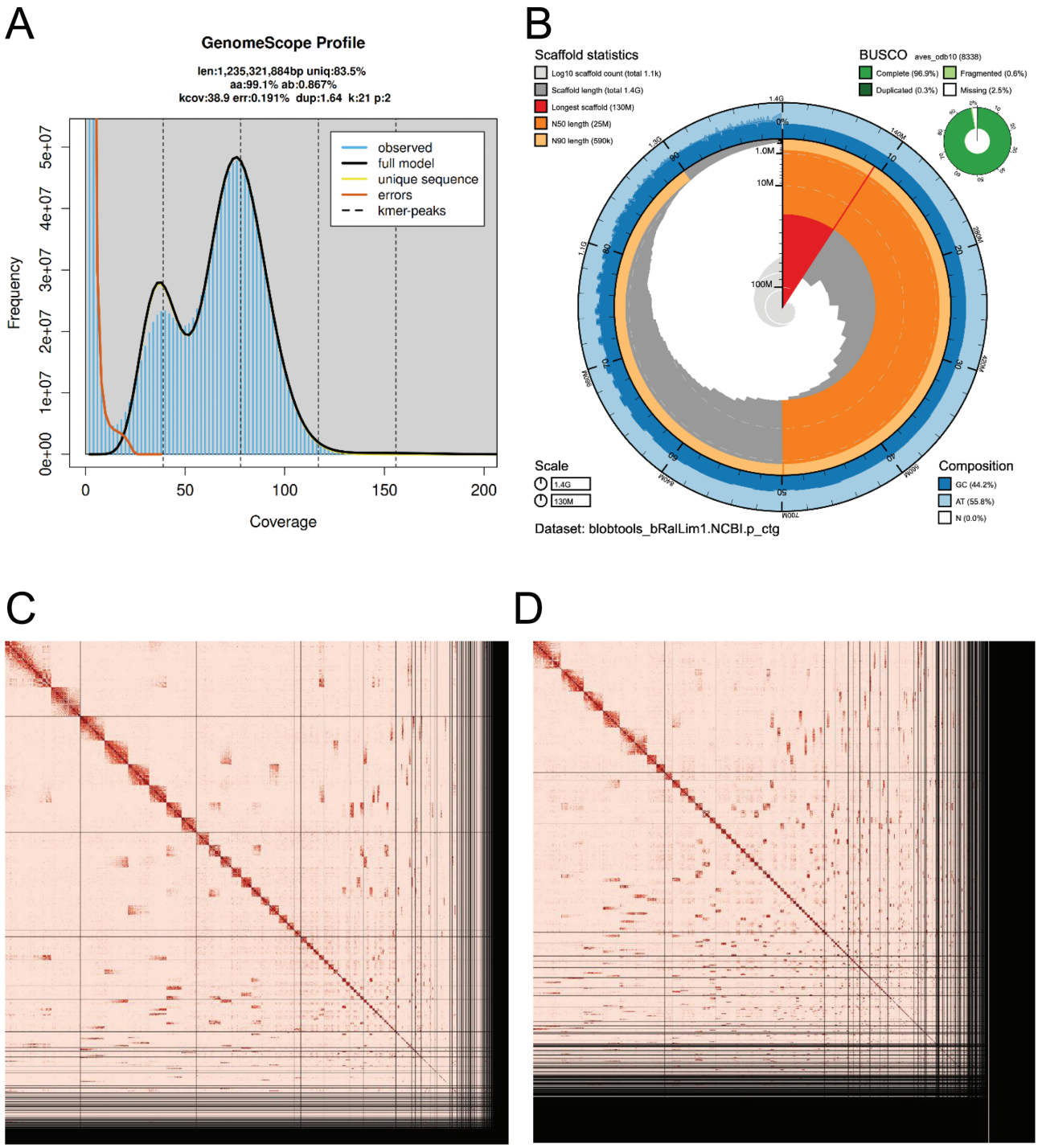

Fig. 2. Visual overview of genome assembly metrics. A) K-mer spectrum generated from PacBio HiFi data without adapters using GenomeScope2.0. The bimodal pattern observed corresponds to a diploid genome. K-mers covered at lower coverage and lower frequency correspond to differences between haplotypes, 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 Virginia rail primary assembly (bRalLim1.0.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 2 of such regions. Black lines differentiate scaffolds.

Table 2. Reference genome assembly metrics for the Virginia rail, *Rallus limicola*, genome assembled by the CCGP.

Bio Projects and Vouchers	**CCGP NCBI BioProject**		**PRJNA720569**			
	Genera NCBI BioProject		PRJNA765853			
	Species NCBI BioProject		PRJNA777214			
	NCBI BioSample		SAMN24505263			
	Specimen identification		171378401			
	NCBI Genome accessions		Primary		Alternate	
	Assembly accession		JAKCOZ000000000		JAKCPA000000000	
	Genome sequences		GCA_022605955.1		GCA_022605895.1	
Genome Sequence	PacBio HiFi reads	Run	1 PACBIO_SMRT (Sequel II) run: 3.7 M spots, 50.2 G bases, 11.8 Gb			
		Accession	SRX14637562			
	Omni-C Illumina reads	Run	1 ILLUMINA (Illumina NovaSeq 6000) run: 184.5M spots, 55.7G bases, 18Gb			
		Accession	SRX14637563, SRX14637564			
Genome Assembly Quality Metrics	Assembly identifier (Quality code[a])		bRalLim1(7.7.Q61.C73)			
	HiFi Read coverage[b]		40.59X			
			Primary		Alternate	
	Number of contigs		1,299		5,118	
	Contig N50 (bp)		11,906,378		2,459,492	
	Contig NG50 (bp)[b]		14,683,902		3,106,228	
	Longest Contigs		45,782,814		20,169,545	
	Number of scaffolds		1,102		4,450	
	Scaffold N50 (bp)		25,302,669		9,049,408	
	Scaffold NG50 (bp)[b]		30,269,380		14,683,902	
	Largest scaffold		128,426,625		79,253,198	
	Size of final assembly (bp)		1,395,598,026		1,425,697,905	
	Gaps per Gbp (#Gaps)		141 (197)		468 (668)	
	Indel QV (Frame shift)		40.92047449		40.97515036	
	Base pair QV		61.2684		62.2928	
					Full assembly = 61.7558	
	k-mer completeness		89.1228		89.0825	
					Full assembly = 99.5675	
	BUSCO completeness (Aves) n= 8338		C	S	D	F	M
		P[c]	96.90%	96.60%	0.30%	0.60%	2.50%
		A[c]	96.50%	95.90%	0.60%	0.70%	2.80%
	Organelles		1 Complete mitochondrial sequence CM040152				

[a]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 2n = 76 inferred from other species in the same family, *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.
[b]Read coverage and NGx statistics have been calculated based on the estimated genome size of 1.2 Gb.
[c]P(rimary) and (A)lternate assembly values.

similar to the estimated value from Genomescope 2.0 (Figure 2A). The primary assembly is more contiguous and consists of 1,102 scaffolds spanning 1.39 Gb with a contig N50 of 11.0 Mb, scaffold N50 of 25.3 Mb, largest contig of 45 Mb, and largest scaffold of 128.4 Mb. In contrast, the alternate assembly consists of 4,450 scaffolds, spanning 1.42 Gb with a contig N50 of 2.4 Mb, scaffold N50 of 9.04 Mb, largest contig of 20.1 Mb, and largest scaffold of 79.2 Mb. Detailed assembly metrics are reported in Table 2 and Figure 2B. The curation process for the primary assembly required us to break 5 of the joins generated during scaffolding corresponding to misassemblies. We closed 3 gaps on the primary assembly and 10 on the alternate. Based on NCBI feedback, we trimmed 3 remainder sequencing adapters (43, 43 and 23 bp long). Contact maps for the assemblies presented here show some level of fragmentation, but also little evidence of inversions and translocations within contigs (Figure 2C and D). The primary assembly has a BUSCO completeness score of 96.9% using the Aves gene set, a per base quality (QV) of 61.25, a k-mer completeness of 89.12, and a frameshift indel QV of 40.92. The alternate assembly has a BUSCO completeness score of 96.5% using the Aves gene set, a per base quality (QV) of 62.29, a k-mer completeness of 89.08, and a frameshift indel QV of 40.97. 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 is18,921 bp in length, with a final base composition A = 32.514%, C = 30.352%, G = 12.034%, and T = 25.1%. The final assembly consists of 22 unique transfer RNAs and 13 protein coding genes.

Discussion

The genome assembly presented here for the Virginia rail is 1 of 7 publicly available for the family Rallidae and the only genome assembly available in the genus *Rallus*. Compared with other species of Rallidae, the 1.39 Gb Virginia rail genome was similar in size (mean = 1.2 Gb, range = 1.1 to 1.4 Gb), with above average contig N50 of 11.9 Mb (Rallidae mean = 4.7 Mb, range = 0.01 to 13.5 Mb), and scaffold N50 of 25.3 Mb (Rallidae mean = 20.8 Mb, range = 0.04 to 71.6 Mb). The GC composition of the Virginia rail genome (44%) was similar to the average GC composition for Rallidae (mean = 43%, range = 43% to 44%).

The Virginia rail genome assembly will serve as an important resource to improve understanding of rail phylogeography, biology, and ecology. With more than 40 genera, the Rallidae is a diverse and globally distributed family with a complex evolutionary history (Kirchman 2012; Garcia-R and Matzke 2021; Kirchman et al. 2021). For example, Tavares et al. (2010) suggested that limited gene flow among disjunct wetlands across the distribution of water rails (*R. aquaticus*) in Asia and Europe has resulted in the formation of a distinct species of water rail breeding in East Asia (*R. indicus*). Similar studies of gene flow across the broad distribution of the Virginia rail in North America could help resolve phylogeographic relationships in Virginia rails. Such studies may also highlight differences in the landscape genetic architecture of migratory and nonmigratory populations of Virginia rails, and perhaps offer insight into the genetic basis of this important life history variation. In addition, future studies will aid conservation efforts by using this genome to infer Virginia rail movements, estimate effective population sizes, and assess the potential of rails for adaptive evolution in the face of growing threats, including climate change and wetland loss.

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 thank E. Burkett, J. Garcia, and C. Hirano for assistance with permits, and Sean 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 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.

Data availability

Data generated for this study are available under NCBI BioProject PRJNA795167. Raw sequencing data for this sample (NCBI BioSample SAMN24505263) are deposited in the NCBI Short Read Archive (SRA) under SRX14637562 for PacBio HiFi data and SRX14637563 and SRX14637564 for Omni-C Illumina 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. https://doi.org/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. https://doi.org/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 nonequilibrium metapopulation dynamics driven by biogeography. *Ecol Lett*. 2022:25(11):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. https://doi.org/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;10:1361–1374. https://doi.org/10.1534/g3.119.400908

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

Conway CJ. Virginia rail (*Rallus limicola*), version 1.0. In: Poole AE, Gill FB, editors. *Birds of the world*. Ithaca (NY): 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. https://doi.org/10.1093/jhered/esac046

Fournier AMV, Mengel DC, Krementz DG. Virginia and Yellow Rail autumn migration ecology: synthesis using multiple data sets. *Anim Migr*. 2017a;4(1):15–22. https://doi.org/10.1515/ami-2017-0003

Fournier AMV, Sullivan AR, Bump JK, Perkins M, Shieldcastle MC, King S. L. Combining citizen science species distribution models and stable isotopes reveals migratory connectivity in the secretive

Virginia rail. *J Appl Ecol*. 2017b;54(2):618–627. https://doi.org/10.1111/1365-2664.12723

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. https://doi.org/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. https://doi.org/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. *PLOS Computational Biology*. 2019;15(8):e1007273. https://doi.org/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. https://doi.org/10.1007/s10592-010-0126-4

Goloborodko A, Abdennur N, Venev S, Brandao HB, Fudenberg G. *mirnylab/pairtools*: *v0.2.0 (v0.2.0)*. *Zenodo*. 2018. https://doi.org/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. https://doi.org/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. https://doi.org/10.1093/bioinformatics/btt086

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. https://doi.org/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. https://doi.org/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):1–21. https://doi.org/10.1093/ornithology/ukab042

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. https://doi.org/10.1093/gigascience/gix085

Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv, arXiv: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. arXiv, arXiv:2106.11799 [q-bio], 2021, preprint: not peer reviewed.

Ramírez F, Bhardwaj V, Arrigoni L, Lam KC, Grüning BA, Villaveces J, Manke T. High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat Commun*. 2018;9(1). https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1186/s13059-020-02134-9

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. https://doi.org/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 Genomics*. 2022;23:157. https://doi.org/10.1186/s12864-022-08375-1

Tavares ES, De Kroon GH, Baker AJ. Phylogenetic and coalescent analysis of three loci suggest that the Water Rail is divisible into two species, *Rallus aquaticus* and *R. indicus*. *BMC Evol Biol*. 2010;10(1):226. https://doi.org/10.1186/1471-2148-10-226