A Highly Contiguous Reference Genome for Northern Bobwhite (Colinus virginianus)

Jessie F. Salter,* Oscar Johnson,* Norman J. Stafford, III,† William F. Herrin, Jr.,‡ Darren Schilling,§ Cody Cedotal,** Robb T. Brumfield,* and Brant C. Faircloth*†

*Museum of Natural Science and Department of Biological Sciences, Louisiana State University, Baton Rouge, Louisiana, †39731 LA-10, Franklinton, Louisiana, ‡15531 Chanove Avenue, Greenwell Springs, Louisiana, §Franklinton Veterinary Clinic, 525 11th Avenue, Franklinton, Louisiana, and **Louisiana Department of Wildlife and Fisheries, P.O. Box 98000, Baton Rouge, Louisiana

ABSTRACT Northern bobwhites (Colinus virginianus) are small quails in the New World Quail family (Odontophoridae) and are one of the most phenotypically diverse avian species. Despite extensive research on bobwhite ecology, genomic studies investigating the evolution of phenotypic diversity in this species are lacking. Here, we present a new, highly contiguous assembly for bobwhites using tissue samples from a vouchedered, wild, female bird collected in Louisiana. By performing a de novo assembly and scaffolding the assembly with Dovetail Chicago and HiC libraries and the HiRise pipeline, we produced an 866.8 Mb assembly including 1,512 scaffolds with a scaffold N50 of 66.8 Mb, a scaffold L90 of 17, and a BUSCO completeness score of 90.8%. This new assembly represents approximately 96% of the non-repetitive and 84% of the entire bobwhite genome size, greatly improves scaffold lengths and contiguity compared to an existing draft bobwhite genome, and provides an important tool for future studies of evolutionary and functional genomics in bobwhites.

INTRODUCTION Northern bobwhites (Colinus virginianus; hereafter bobwhites) are widely distributed quails primarily found in pine woodlands and grasslands of the eastern United States and Mexico (Brennan 1999). Bobwhites hold a significant place in the cultural heritage of both countries due to their status as popular game birds (Bent 1963; Burger et al. 1999), and they have also played a significant role in biological research because they are one of the most intensively studied birds in the world (Guthery 1997). Bobwhites are remarkably polytypic: there are 22 subspecies recognized by male plumage (Brennan 1999) – a larger number of subspecies than 99% of all other birds (Dickinson and Remsen 2013). Although bobwhite ecology research has been extensive, the evolutionary relationships between bobwhite subspecies remain murky (Ellsworth et al. 1989; Evans et al. 2009; Eo et al. 2009; Williford 2013; Williford et al. 2014, 2016) and the genetic basis of phenotypic diversity in bobwhites has been largely unstudied (but see Cole et al. 1949).

Identifying genotypes associated with specific phenotypes increasingly relies on whole genome sequencing, particularly for investigating the genetic basis of phenotypic differences in non-model organisms (Ellegren 2014). The first draft genome assembly for bobwhites (GCA_000599485.1; hereafter Cv_TX_1.1) was generated from an unvouchered wild female bird from Texas (Halley et al. 2014). Cv_TX_1.1 used small and medium insert paired-end (PE) and mate pair (MP) libraries to produce a 1.172 Gb genome assembly with 77x coverage, 50% of the assembly in scaffolds of at least 45.5 Kbp in size (N50), and 90% of the assembly in 25,837 scaffolds (L90, Halley et al. 2014). Sequencing of additional PE and MP libraries from the same bird were used to generate a second assembly (GCA_000599465.2; hereafter Cv_TX_2.0), which yielded a 1.5-fold increase in coverage (122x), a 45-fold improvement in N50 (2.042 Mb), and a nearly threefold decrease in L90 (8,990 scaffolds; Oldeschulte et al. 2017). Although Cv_TX_2.0 was a marked improvement over Cv_TX_1.1, the scaffolds remained relatively short, which can hinder identification of structural variants (Domyan et al. 2014). Recent studies in

KEYWORDS Colinus virginianus bobwhites Dovetail Chicago HiC genome assembly
birds and other taxa have demonstrated the importance of structural variants in generating morphological diversity within closely-related taxa (Lamichhaney et al. 2016; Tuttle et al. 2016; Vijay et al. 2016), highlighting the need for highly contiguous genome assemblies in phenotype-genotype studies (Wellenreuter and Bernatchez 2018). Here, we describe Cv_LA_1.0, a new assembly for bobwhites using DNA extracted from a vouchered, wild female bird collected in Louisiana. To generate this assembly, we scaffolded contigs from small insert libraries with reads from Chicago (Putnam et al. 2016) and HiC (Lieberman-Aiden et al. 2009) methodologies and the HiRise assembly pipeline (Dovetail Genomics, LLC). The resulting Cv_LA_1.0 assembly is highly contiguous and represents a 32-fold increase in N50 and 528-fold decrease in L90 relative to Cv_TX_2.0 (Oldeschulte et al. 2017).

### METHODS AND MATERIALS

#### Specimen collection and DNA extraction

We collected blood, liver, and other tissues for direct storage in liquid nitrogen from a wild, female bird legally harvested at Sandy Hollow Wildlife Management Area (30.827 N, 90.397 W) in Tangipahoa Parish, Louisiana. After tissue collection, we prepared a specimen for the LSU Museum of Natural Science (LSUMNS) Collection of Birds (LSUMZ 197699), and we stored tissue samples from this specimen in the LSUMNS Collection of Genetic Resources (LSUMZ B-91918). We shipped blood and liver to Dovetail Genomics, LLC (Scotts Valley, CA) where Dovetail Staff performed DNA extraction, library preparation, sequencing, and assembly steps. Dovetail staff extracted high molecular weight (HMW) DNA from tissues using the Blood and Cell Culture Midi Kit (Qiagen, GmbH) following the manufacturer’s protocol. Mean fragment length of the extracted DNA was 85 kb.

#### Short-insert library preparation, sequencing, and assembly

Dovetail staff randomly fragmented extracted DNA by sonication using a Bioruptor Pico (Diagenode, Inc.) and 7 cycles of sonication for 15 sec followed by 90 sec of rest. Dovetail staff then prepared a sequencing library by inputting fragmented DNA to the TruSeq DNA PCR-Free Library Preparation Kit (Illumina, Inc.) following the manufacturer’s protocol. Resulting libraries were sequenced on an Illumina HiSeq X platform using paired-end (PE) 150 bp sequencing. Resulting data were trimmed for low-quality bases and adapter contamination using Trimomatic (Bolger et al. 2014) and used to assemble scaffolds with Meraculous v2.2.5 (Chapman et al. 2011). Before assembly, Dovetail staff used Jellyfish (Marçais and Kingsford 2011) with in-house software similar to GenomeScope (Vurtourke et al. 2017) to profile the short insert reads at a variety of k-mer values (25, 55, 85, 109), estimate genome size, and fit negative binomial models to the data. The resulting profiles suggested a k-mer size of 55 was optimal for assembly, and Dovetail staff assembled contigs using Meraculous with a k-mer size of 55, a minimum k-mer frequency of 12, and the diploid nonredundant haplotigs mode.

#### Chicago library preparation and sequencing

Following de novo assembly with Meraculous, Dovetail staff prepared a single, proprietary “Chicago” library following the methods described in Putnam et al. (2016). Briefly, they reconstituted ~500 ng of HMW genomic DNA into chromatin in vitro and fixed the reconstituted DNA with formaldehyde. Then, they digested fixed chromatin with DpnII, filled in 5’ overhangs with biotinylated nucleotides, and ligated free blunt ends. After ligation, they reversed crosslinks and purified the DNA from protein. Dovetail staff treated purified DNA to remove biotin that was not internal to ligated fragments and sheared the resulting DNA to ~350 bp mean fragment size using a Bioruptor Pico. Dovetail staff then prepared sequencing libraries from the sheared DNA using NEBNext Ultra enzymes (New England Biolabs, Inc.) and Illumina-compatible adapters. They isolated biotin-containing fragments using streptavidin beads before PCR enrichment of each library.

Dovetail staff then sequenced amplified libraries on an Illumina HiSeq X platform using PE 150 reads to approximately 90X depth.

#### Dovetail HiC library preparation and sequencing (multiple libraries)

Dovetail staff also prepared two Dovetail HiC libraries following the procedures outlined in Lieberman-Aiden et al. (2009). Briefly, for each library, Dovetail staff used formaldehyde to fix chromatin in place in the nucleus. They extracted and digested fixed chromatin with DpnII, filled in the 5’ overhangs with biotinylated nucleotides, and ligated free blunt ends. After ligation, Dovetail staff reversed crosslinks and purified the DNA from protein. They treated the purified DNA to remove biotin that was not internal to ligated fragments and sheared the DNA to ~350 bp mean fragment size using a Bioruptor Pico. Dovetail staff then prepared sequencing libraries using NEBNext Ultra enzymes and Illumina-compatible adapters. They isolated biotin-containing fragments using streptavidin beads before PCR enrichment of each library and sequenced the resulting libraries on an Illumina HiSeq X Platform using PE 150 reads to approximately 60X depth.

#### Assembly scaffolding with HiRise

To scaffold and improve the bobwhite assembly, Dovetail staff input the de novo assembly from Meraculous, along with shotgun reads, Chicago library reads, and Dovetail HiC library reads into HiRise (April 2017 version), a software pipeline designed for this purpose (Putnam et al. 2016). Using HiRise, Dovetail staff conducted an iterative analysis. First, they aligned shotgun and Chicago library sequences to the draft contig assembly using a modified SNAP read mapper (http://snap.cs.berkeley.edu). Second, they analyzed the separations of Chicago read pairs mapped within draft scaffolds to produce a likelihood model for genomic distance between read pairs, and they used this model to: identify and break putative misjoins, score prospective joins, and make joins above a threshold. Finally, after aligning and scaffolding the draft assembly using the Chicago data, Dovetail staff aligned and scaffolded the Chicago assembly using Dovetail HiC library sequences following the same method. After scaffolding, Dovetail staff...
used the short-insert sequences to close remaining gaps between contigs where possible.

**Assembly polishing, contiguity statistics, and BUSCO analyses**

After receiving the assembly from Dovetail, we aligned the short insert data back to the scaffolded assembly using bow v0.7.17-r1188 (Li and Durbin 2009) and samtools v1.9 (Li et al. 2009) and polished the scaffolds using Pilon v1.23 (Walker et al. 2014) on a 48-core, 1.5 TB RAM compute node with default parameters. After polishing, we computed contiguity statistics of our scaffolded assembly as well as the Cv_TX_2.0 assembly (Oldeschulte et al. 2017) using QUAST v5.0.2 (Mikheenko et al. 2018), UCSC Browser Utilities (Kent et al. 2002), and GNU Coreutils (https://www.gnu.org/software/coreutils/), and we performed BUSCO analyses against both genomes using BUSCO v3.1.0 (Waterhouse et al. 2018) and the Aves Data Set (aves_odb9).

**Data availability**

Data from all sequencing runs and the final assembly, Cv_LA_1.0, are available from NCBI BioProject (PRJNA454855). Short-insert, Chicago, and HiC reads are also available from the NCBI SRA (SRP215501), and the assembly is available from NCBI Genome using the accession VONY0000000. The version described in this manuscript is VONY01000000. Outputs from QUAST and BUSCO analyses are available as supplemental files from figshare. Supplemental material available at figshare: https://doi.org/10.25387/g3.9273542.

**RESULTS AND DISCUSSION**

Sequencing of short-insert libraries produced 441.8 million read pairs with an average insert size of 428 bp. Analysis of the k-mer histogram at the optimal value of 55 suggested the genome size was 1.0 Gb, and the estimated Q20 read depth for this genome size was approximately 118x. Meraculous assembly using a k-mer value of 55 produced 23,275 contigs having a total length of 853.1 Mb and an N50 of 113.6 Kb. These contigs were joined by Meraculous into 14,482 scaffolds totaling 854.1 Mb in length with an N50 of 15.5 Mb and a L90 of 53 scaffolds. The longest Meraculous scaffold was 1.6 Mb. Meraculous estimated that the assembled contigs comprised 96% of the estimated, non-repetitive genome size and 84% of the entire genome size.

Chicago library sequencing produced 303 million read pairs, and the estimated physical coverage (the number of read pairs with inserts between 10 and 10,000 kb) spanning each position in the Chicago assembly was 38,615. HiRise made 147 joins and zero breaks to the Chicago-scaffolded assembly to produce a HiC assembly including 1,512 scaffolds totaling 866.8 Mb in length with an N50 of 66.9 Mb and a L90 of 17 scaffolds. The longest HiC scaffold was 180.8 Mb.

After polishing the HiC assembly, the bobwhite genome assembly Cv_LA_1.0 included 1,512 scaffolds having an N50 of 66.8 Mb and a L90 of 17. Comparison of Cv_LA_1.0 with the Cv_TX_2.0 assembly (Table 1) shows the increase in contiguity of our assembly relative to the assembly produced by Oldeschulte et al. (2017). BUSCO analyses of both genomes are similar (Table 2), although we detected slightly fewer BUSCOs (-0.7%) in our Cv_LA_1.0 assembly relative to Cv_TX_2.0, perhaps due to repeat regions that were excluded from the contigs assembled by Meraculous. Future improvements to this assembly will incorporate Pacific Biosciences long-read sequences to help fill gaps that are likely associated with repeat regions that were difficult to assemble using short-reads.

**ACKNOWLEDGMENTS**

We thank Shaune Hall and other staff members at Dovetail Genomics for working with us, and Donna Dittmann and Steve Cardinali for assistance with specimen preparation and tissue storage. Special thanks to Chick and Lulu for assistance in the field. Funding for this project was provided by Louisiana State University, National Science Foundation grants DEB-1242260 and IOS-1754417 (to BCF) and DEB-1655624 (to BCF and RTB). JFS was supported by the LSU Museum of Natural Science, and portions of this research were conducted with high performance computational resources provided by the Louisiana Optical Network Infrastructure (http://www.loni.org). The individual bobwhite used in this study was collected by a private quail hunter with a valid Louisiana hunting license. NJS, WFH, DS, CC, JFS, and OJ performed field work; and JFS and OJ prepared specimens; BCF performed analyses; JFS and BCF wrote the paper; BCF and RTB provided funding; and all authors reviewed and approved the manuscript.

**LITERATURE CITED**

Bent, A. C., 1963 Life Histories of North American Gallinaceous Birds, Dover Publications, Inc., New York.

Bolger, A. M., L. Mohse, and B. Usadel, 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Brennan, L. A., 1999 Northern bobwhite (Colinus virginianus). The birds of North America, number 397. National Academy of Natural Sciences, Philadelphia, Pennsylvania, and The American Ornithologists’ Union, Washington, D.C.

Burger, L. W., D. A. Miller, and R. I. Southwick, 1999 Economic impact of northern bobwhite hunting in the southeastern United States. Wildl. Soc. Bull. 27: 1010–1018.
Chapman, J. A., I. Ho, S. Sunkara, S. Luo, G. P. Schroth et al., 2011 Meraculous: de novo genome assembly with short paired-end reads. PLoS One 6: e23501. https://doi.org/10.1371/journal.pone.0023501

Cole, L. J., H. L. Stoddard, and E. V. Komarek, 1949 Red bob-white: A report and correction. Auk 66: 28–35. https://doi.org/10.2307/4080656

Dickinson, E. C., and J. V. Remsen, Jr. (Editors), 2013 The Howard & Moore Complete Checklist of the Birds of the World, Aves Press, Eastbourne, U.K.

Domyan, E. T., M. W. Guerney, Z. Kronenberg, S. Krishnan, R. E. Boissy et al., 2014 Epistatic and combinatorial effects of pigmentary gene mutations in the domestic pigeon. Curr. Biol. 24: 459–464. https://doi.org/10.1016/j.cub.2014.01.020

Ellegren, H., 2014 Genome sequencing and population genomics in non-model organisms. Trends Ecol. Evol. 29: 51–63. https://doi.org/10.1016/j.tree.2013.09.008

Ellsworth, D. L., J. L. Roseberry, and W. D. Klimstra, 1989 Genetic structure and gene flow in the northern bobwhite. Auk 106: 492–495.

Eo, S. H., J. P. Wares, and J. P. Carroll, 2009 Subspecies and units for conservation and units for the conservation of the northern bobwhite in the eastern United States. Conserv. Genet. 11: 867–875. https://doi.org/10.1007/s10592-009-9926-9

Evans, K. O., M. D. Smith, L. W. Burger, R. J. Chambers, A. E. Houston et al., 2009 Release of pen-reared bobwhites: Potential consequences to the genetic integrity of resident wild populations National Quail Symposium Proceedings 6: 121–133.

Guthery, F. S., 1997 A philosophy of habitat management for northern bobwhites. J. Wildl. Manage. 61: 291–301. https://doi.org/10.2307/3802584

Halley, Y. A., S. E. Dowd, J. E. Decker, P. M. Seabury, E. Bhattarai et al., 2014 A draft de novo genome assembly for the northern bobwhite (Colinus virginianus) reveals evidence for a rapid decline in effective population size beginning in the Late Pleistocene. PLoS One 9: e92040. https://doi.org/10.1371/journal.pone.0092040

Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle et al., 2002 The human genome browser at UCSC. Genome Res. 12: 996–1006. https://doi.org/10.1101/gr.229102

Lamichhaney, S., G. Fan, F. Widemo, U. Gunnarsson, D. S. Thalmann et al., 2016 Structural genomic changes underlie alternative reproductive strategies in the ruff (Philomachus pugnax). Nat. Genet. 48: 84–88. https://doi.org/10.1038/ng.3430

Li, H., and R. Durbin, 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Lieberman-Aiden, E., N. I. van Berkum, L. Williams, M. Imakaev, T. Ragoczy et al., 2009 Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326: 289–293. https://doi.org/10.1126/science.1181369

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan et al., 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Marçais, G., and C. Kingsford, 2011 A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27: 764–770. https://doi.org/10.1093/bioinformatics/btr011

Mikheenko, A., A. Pribelski, V. Saveliev, D. Antipov, and A. Gurevich, 2018 Versatile genome assembly evaluation with QUAST-LG. Bioinformatics 34: i142–i150. https://doi.org/10.1093/bioinformatics/bty266

Oldeschulte, D. L., Y. A. Halley, M. L. Wilson, E. K. Bhattarai, W. Brasher et al., 2017 Annotated draft genome assemblies for the northern bob-white (Colinus virginianus) and the scaled quail (Callipepla squamata) reveal disparate estimates of modern genome diversity and historic effective population size. G3 (Bethesda) 7: 3047–3058. https://doi.org/10.1534/g3.117.034083

Putnam, N. H., B. L. O’Connell, J. C. Stites, B. J. Rice, M. Blanchette et al., 2016 Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome Res. 26: 342–350. https://doi.org/10.1101/gr.193474.115

Tuttle, E. M., A. O. Bergland, M. L. Korody, M. S. Brewer, D. J. Newhouse et al., 2016 Divergence and functional degradation of a sex chromosome-like supergene. Curr. Biol. 26: 344–350. https://doi.org/10.1016/j.cub.2015.11.069

Vijay, N., C. M. Bossu, J. W. Poelstra, M. H. Weinsteinester, A. Suh et al., 2016 Evolution of heterogeneous genome differentiation across multiple contact zones in a crow species complex. Nat. Commun. 7: 13195. https://doi.org/10.1038/ncomms13195

Vurture, G. W., F. J. Sedlacek, M. Nattestad, C. J. Underwood, H. Fang et al., 2017 GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics 33: 2202–2204. https://doi.org/10.1093/bioinformatics/btx153

Walker, B. J., T. Abeel, T. Shea, M. Priest, A. Abouelliel et al., 2014 Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9: e112963. https://doi.org/10.1371/journal.pone.0112963

Waterhouse, R. M., M. Seppey, F. A. Simão, M. Manni, P. Ioannidis et al., 2015 BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol. Biol. Evol. 32: 543–548. https://doi.org/10.1093/molbev/msv339

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

Willford, D. L., 2013 Molecular genetics of the northern bobwhite, scaled quail, and Gambel’s quail. Ph.D. Dissertation, Texas A&M University-Kingsville, Kingsville.

Willford, D., R. W. Deyoung, R. L. Honeycutt, L. A. Brennan, and F. Hernández, 2016 Phylogeography of the bobwhite (Colinus) quails. Wildl. Monogr. 193: 1–176. https://doi.org/10.2307/wi1931

Willford, D., R. W. Deyoung, R. L. Honeycutt, L. A. Brennan, H. F. Wehland et al., 2014 Contemporary genetic structure of the northern bobwhite west of the Mississippi River. J. Wildl. Manage. 78: 914–929. https://doi.org/10.1002/jwmg.733

Communicating editor: A. Sethuraman