G3 Genes | Genomes | Genetics

GENOME REPORT

De Novo Assembly of a High-Quality Reference Genome for the Horned Lark (*Eremophila alpestris*)

Nicholas A. Mason,*,†,‡,1 Paulo Pulgarin,§, Carlos Daniel Cadena,§ and Irby J. Lovette*,†**

*Fuller Evolutionary Biology Program, Cornell Lab of Ornithology, †Department of Ecology and Evolutionary Biology, Cornell University, Ithaca, New York 14853, ‡Museum of Vertebrate Zoology, University of California, Berkeley, California, 94720, §Laboratorio de Biología Evolutiva de Vertebrados, Departamento de Ciencias Biológicas, Universidad de Los Andes, Bogotá, Colombia, and **Facultad de Ciencias y Biotecnología, Universidad CES, Medellín, Colombia

ORCID IDs: 0000-0002-5266-463X (N.A.M.); 0000-0003-4937-2529 (P.P.); 0000-0003-4530-2478 (C.D.C.); 0000-0003-0268-4875 (I.J.L.)

ABSTRACT The Horned Lark (*Eremophila alpestris*) is a small songbird that exhibits remarkable geographic variation in appearance and habitat across an expansive distribution. While *E. alpestris* has been the focus of many ecological and evolutionary studies, we still lack a highly contiguous genome assembly for the Horned Lark and related taxa (Alaudidae). Here, we present CLO_EAlp_1.0, a highly contiguous assembly for *E. alpestris* generated from a blood sample of a wild, male bird captured in the Altiplano Cundiboyacense of Colombia. By combining short-insert and mate-pair libraries with the ALLPATHS-LG genome assembly pipeline, we generated a 1.04 Gb assembly comprised of 2713 scaffolds, with a largest scaffold size of 31.81 Mb, a scaffold N50 of 9.42 Mb, and a scaffold L50 of 30. These scaffolds were assembled from 23685 contigs, with a largest contig size of 1.69 Mb, a contig N50 of 193.81 kb, and a contig L50 of 1429. Our assembly pipeline also produced a single mitochondrial DNA contig of 14.00 kb. After polishing the genome, we identified 94.5% of single-copy gene orthologs from an Aves data set and 97.7% of single-copy gene orthologs from a vertebrata data set, which further demonstrates the high quality of our assembly. We anticipate that this genomic resource will be useful to the broader ornithological community and those interested in studying the evolutionary history and ecological interactions of larks, which comprise a widespread, yet understudied lineage of songbirds.

KEYWORDS
Alaudidae
ALLPATHS-LG
Eremophila alpestris
genome assembly
horned lark

The Horned Lark (*Eremophila alpestris*) is a widespread species of songbird that occupies grasslands, tundras, deserts, and other sparsely vegetated habitats on five continents (Beason 1995). As is characteristic of most species in the family Alaudidae, *E. alpestris* is a terrestrial species that nests on the ground and relies on camouflage to avoid predation by avian predators (Donald *et al.* 2017). The Horned Lark has been studied extensively in terms of geographic variation and systematics (Behle 1942; Johnson 1972), population genetics (Drovetski *et al.* 2006, 2014; Mason *et al.* 2014; Ghorbani *et al.* 2019), physiological adaptations (Trost 1972), breeding biology (de Zwaan *et al.* 2019), and responses to human activity, such as agriculture (Mason and Unitt 2018) and wind energy (Erickson *et al.* 2014), among other focal areas. Despite extensive past and ongoing research involving *E. alpestris* and other alaudids, we lack a highly contiguous reference genome for the species and the family as a whole (but see Dierickx *et al.* 2019). Generating genomic resources for the Horned Lark and related taxa will enable studies linking phenotypic and genetic variation (Kratochwil and Meyer 2015; Hoban *et al.* 2016), chromosomal rearrangements (Wellenreuther and Bernatchez 2018), and many other avenues of future genomic research for non-model organisms (Ellegren 2014).

Here, we describe CLO_EAlp_1.0, a new genomic assembly that we built with DNA extracted from a wild, male lark captured from a demographically small and geographically isolated population near Toca, Boyacá, Colombia. We sampled this individual and population because it had high *a priori* likelihood of high homozygosity compared to larks elsewhere with much larger effective population sizes and variable patterns of connectivity to adjacent populations. To generate this *de novo* assembly, we used the ALLPATHS-LG pipeline (Butler *et al.* 2008;

Copyright © 2020 Mason *et al.*
doi: https://doi.org/10.1534/g3.119.400846
Manuscript received October 16, 2019; accepted for publication December 18, 2019; published Early Online December 18, 2019.
This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Supplemental material available at figshare: https://doi.org/10.25387/g3.9992360.
[1]Corresponding author: Museum of Vertebrate Zoology and Department of Integrative Biology, University of California, 3101 Valley Life Science Building, Berkeley, CA 94720. E-mail: nmason@berkeley.edu

Gnerre *et al.* 2011). Given the lack of genomic resources currently available for Alaudidae, we hope this *de novo* assembly will inspire and facilitate future studies on the genomic biology of larks—a widespread, diverse lineage of songbirds.

MATERIALS AND METHODS

Sample collection, DNA extraction, and sequencing

We captured a male *E. alpestris* (EALPPER07; NCBI BioSample SAMN12913182) approximately 170 km NE of Bogotá, Colombia near the town of Toca on the shores of the Embalse de La Copa in the Altiplano Cundiboyacense of the Boyacá department (5.623299, -73.184156). This population is small and represents a subspecies (*E. a. peregrina*) that is geographically isolated from other populations of larks, the nearest population of which is in Oaxaca, Mexico. The Colombian subspecies of Horned Lark likely underwent a population bottleneck upon colonizing the distant, high-elevation plateaus of the Altiplano Cundiboyacense region and therefore probably has high homozygosity compared to other populations, which is preferable for *de novo* genome assembly. We collected blood from the brachial vein, from which we subsequently extracted genomic DNA with a Gentra Puregene Blood Kit (Qiagen, Hilden, Germany) following the manufacturer's protocol. We confirmed the sex of the individual using PCR amplification (Chu *et al.* 2015). After running the sample on a 1% agarose gel to confirm the presence of high molecular weight DNA, we sent the extraction to the Cornell Weil Medical School (New York, USA), where they generated a 180 bp fragment library, a 3 kb mate-pair library and a 8 kb mate-pair library. We sequenced the 180 bp library across two lanes and combined the 3 kb and 8 kb mate-pair libraries on another lane of Illumina HiSeq 2500 to perform 100 bp paired-end sequencing.

Genome assembly, polishing, and assessment

We assembled the genome with ALLPATHS-LG v52415 (Butler *et al.* 2008; Gnerre *et al.* 2011). We did not perform additional adapter removal or quality filtering with the short-insert 180 bp libraries because ALLPATHS-LG has built-in steps that remove low-quality and adapter-contaminated reads (Butler *et al.* 2008). Once the initial assembly had finished, we aligned the short-insert and mate-pair libraries back to the assembly genome using bwa 0.7.17-r1188 (Li and Durbin 2009) and samtools v1.9 (Li *et al.* 2009) and then performed three iterations of scaffold polishing using pilon v1.22 (Walker *et al.* 2014) with default parameters. Once scaffold polishing had finished, we ordered and correspondingly renamed the scaffolds with respect to decreasing scaffold size using SeqKit v0.7.2 (Shen *et al.* 2016). We assessed the contiguity of the *de novo* genome using the function stats.sh from BBMap v38.73 (Bushnell 2014) and estimated genome completeness with BUSCO v3 (Simão *et al.* 2015; Waterhouse *et al.* 2018) alongside HMMER v3.1b2 (Finn *et al.* 2011) and BLAST+ v2.7.1 (Camacho *et al.* 2009) to identify single-copy orthologous gene sets among birds and vertebrates. We subsequently submitted our genome to the NCBI genome submission portal, which performs an additional scan for contaminants, including adapter contamination, and removed any additional contaminant sequences that were detected.

Mitochondrial genome assembly

We also assembled the mitochondrial genome for the same individual (EALPPER07) with NOVOplasty v3.7 (Dierckxsens *et al.* 2017) using a ND2 sequence (GenBank Accession KF743558) from a previous study (Mason *et al.* 2014) as the initial seed to begin the assembly process.

Table 1 *De novo* genome assembly metrics estimated using BBMap

Assembly Statistic	CLO_EAlp_1.0
# scaffolds / contigs	2713 / 23684
Largest scaffold / contig	31.81 Mb / 1.69 Mb
Total length	1.04 Gb
Scaffold / contig N50	9.42 Mb / 193.81 kb
Scaffold / contig N90	1.20 Mb / 31.02 kb
Scaffold / contig L50	30 / 1429
Scaffold / contig L90	141 / 6205
# N's per 100 kbp	3472.35
GC (%)	42.23

Data availability

Raw output from sequencing runs and the final assembly, CLO_EAlp_1.0, are available from NCBI (BioProject PRJNA575884). Short-fragment and mate-pair libraries are also available from the NCBI SRA (SUB6392689). Outputs from BUSCO and BBMap analyses are available from figshare (https://doi:10.6084/m9.figshare.9956063; https://doi:10.6084/m9.figshare.9956042). Supplemental material available at figshare: https://doi.org/10.25387/g3.9992360.

RESULTS AND DISCUSSION

Taken together, the three lanes of Illumina HiSeq 2500 sequencing generated 1.59×10^9 total reads (~134x estimated coverage of a 1.2 Gb genome), including 5.45×10^8 paired-end reads for the 180 bp short-insert libraries, 1.24×10^8 paired-end reads for the 3 kb mate-pair library, and 1.27×10^8 paired-end reads for the 8 kb mate-pair library. Following scaffold polishing, the finalized CLO_EAlp_1.0 assembly consisted of 2713 scaffolds that totaled 1.04 Gb. The largest scaffold was 31.81 Mb while the scaffold N50 was 9.42 Mb and scaffold L50 was 30 (Table 1). The assembly consisted of 23,684 contigs, including a largest contig size of 1.69 Mb, a contig N50 of 193.81 kb, and a contig L50 of 1429. The average GC content of the assembly was 42.23%, which is similar to other birds (Jarvis *et al.* 2014; Botero-Castro *et al.* 2017), while the *de novo* genome assembly included 94.5% of single-copy orthologs from the Aves data set and 97.7% of the Vertebrata data set as identified by BUSCO (Table 2). The mitochondrial assembly pipeline generated a single mtDNA contig of 14 kb.

We opted not to assemble pseudochromosomes by aligning our *de novo* genome to an existing chromosome-level genome assembly (*e.g.*, Zebra Finch (*Taeniopygia guttata*). While birds generally exhibit strong synteny (Derjusheva *et al.* 2004), avian sex chromosomes and microchromosomes are often comprised of extensive rearrangements (Volker *et al.* 2010). Our assembly could be further improved such that scaffolds match full chromosomes through strategies such as Hi-C (Burton *et al.* 2013) or ultra-long read sequencing technology

Table 2 Output from BUSCO analyses to assess genome completeness by searching for single-copy orthologs from aves and vertebrata datasets

	Aves	Vertebrata
Complete BUSCOs	4645 (94.5%)	2530 (97.7%)
Complete and single-copy BUSCOs	4590 (93.4%)	2518 (97.4%)
Complete and duplicated BUSCOs	55 (1.1%)	12 (0.5%)
Fragmented BUSCOs	162 (3.3%)	36 (1.4%)
Missing BUSCOs	108 (2.2%)	20 (0.7%)
Total BUSCO groups searched	4915	2586

(Ma *et al.* 2018). Functional annotation of our assembly could also be improved by generating RNA-Seq and protein libraries specifically for larks. Nonetheless, CLO_EAlp_1.0 represents a large step forward toward leveraging the natural history of larks and advanced sequencing technology to further understand avian biology.

ACKNOWLEDGMENTS

We would like to thank Bronwyn Butcher for assistance with lab work. We also thank Leonardo Campagna for helpful advice on genome assembly and bioinformatics. Thanks to Diana Carolina Macana for assistance with field work. This research was supported by SELVA and a Doctoral Dissertation Improvement Grant from the National Science Foundation (DEB-1601072) to NAM.

LITERATURE CITED

Beason, R. C., 1995 Horned Lark (Eremophila alpestris), version 2.0, *The Birds of North America*, Vol. A, edited by Poole, F., and F. B. Gill. Cornell Lab of Ornithology, Ithaca.

Behle, W., 1942 Distribution and variation of the horned larks (*Eremophila alpestris*) of western North America. Univ. Calif. Publ. Zool. 46: 203–316.

Botero-Castro, F., E. Figuet, M.-K. Tilak, B. Nabholz, and N. Galtier, 2017 Avian genomes revisited: hidden genes uncovered and the rates *vs.* traits paradox in birds. Mol. Biol. Evol. 34: 3123–3131. https://doi.org/10.1093/molbev/msx236

Burton, J. N., A. Adey, R. P. Patwardhan, R. Qiu, J. O. Kitzman *et al.*, 2013 Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat. Biotechnol. 31: 1119–1125. https://doi.org/10.1038/nbt.2727

Bushnell, B., 2014 *BBMap: a fast, accurate, splice-aware aligner*, Joint Genome Institute, Walnut Creek, CA.

Butler, J., I. MacCallum, M. Kleber, I. A. Shlyakhter, M. K. Belmonte *et al.*, 2008 ALLPATHS: De novo assembly of whole-genome shotgun microreads. Genome Res. 18: 810–820. https://doi.org/10.1101/gr.7337908

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos *et al.*, 2009 BLAST+: architecture and applications. BMC Bioinformatics 10: 421. https://doi.org/10.1186/1471-2105-10-421

Chu, H., X. Guo, Y. Zeng, X. Zou, S. Guo *et al.*, 2015 A new primer-pair for sex identification of larks and wagtails. Conserv. Genet. Resour. 7: 19–21. https://doi.org/10.1007/s12686-014-0298-1

Derjusheva, S., A. Kurganova, F. Habermann, and E. Gaginskaya, 2004 High chromosome conservation detected by comparative chromosome painting in chicken, pigeon and passerine birds. Chromosome Res. 12: 715–723. https://doi.org/10.1023/B:CHRO.0000045779.50641.00

de Zwaan, D. R., S. Barnes, and K. Martin, 2019 Plumage melanism is linked to male quality, female parental investment and assortative mating in an alpine songbird. Anim. Behav. 156: 41–49. https://doi.org/10.1016/j.anbehav.2019.06.034

Dierckxsens, N., P. Mardulyn, and G. Smits, 2017 NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic Acids Res. 45: e18. https://doi.org/10.1093/nar/gkw955

Dierickx, E., S. Sin, P. van Veelen, M. de L. Brooke, Y. Liu *et al.*, 2019 Neo-sex chromosomes and demography shape genetic diversity in the critically endangered Raso lark. bioRxiv. doi: 10.1101/617563 (Preprint posted April 24, 2019). https://doi.org/10.1101/617563

Donald, P. F., P. Alström, and D. Engelbrecht, 2017 Possible mechanisms of substrate colour-matching in larks (Alaudidae) and their taxonomic implications. Ibis 159: 699–702. https://doi.org/10.1111/ibi.12487

Drovetski, S. V., S. F. Pearson, and S. Rohwer, 2006 Streaked horned lark *Eremophila alpestris strigata* has distinct mitochondrial DNA. Conserv. Genet. 6: 875–883. https://doi.org/10.1007/s10592-005-9074-9

Drovetski, S. V., M. Raković, G. Semenov, I. V. Fadeev, and Y. A. Red'kin, 2014 Limited phylogeographic signal in sex-linked and autosomal loci despite geographically, ecologically, and phenotypically concordant structure of mtDNA variation in the Holarctic avian genus *Eremophila*. PLoS One 9: e87570. https://doi.org/10.1371/journal.pone.0087570

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

Erickson, W. P., M. M. Wolfe, K. J. Bay, D. H. Johnson, and J. L. Gehring, 2014 A comprehensive analysis of small-passerine fatalities from collision with turbines at wind energy facilities. PLoS One 9: e107491. https://doi.org/10.1371/journal.pone.0107491

Finn, R. D., J. Clements, and S. R. Eddy, 2011 HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 39: W29–W37. https://doi.org/10.1093/nar/gkr367

Ghorbani, F., M. Aliabadian, U. Olsson, P. F. Donald, A. A. Khan *et al.*, 2019 Mitochondrial phylogeography of the genus Eremophila confirms underestimated species diversity in the Palearctic. J Ornithol. https://doi.org/10.1007/s10336-019-01714-2

Gnerre, S., I. MacCallum, D. Przybylski, F. J. Ribeiro, J. N. Burton *et al.*, 2011 High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. USA 108: 1513–1518. https://doi.org/10.1073/pnas.1017351108

Hahn, C., L. Bachmann, and B. Chevreux, 2013 Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. Nucleic Acids Res. 41: e129. https://doi.org/10.1093/nar/gkt371

Hoban, S., J. L. Kelley, K. E. Lotterhos, M. F. Antolin, G. Bradburd *et al.*, 2016 Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions. Am. Nat. 188: 379–397. https://doi.org/10.1086/688018

Jarvis, E. D., S. Mirarab, A. J. Aberer, B. Li, P. Houde *et al.*, 2014 Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346: 1320–1331. https://doi.org/10.1126/science.1253451

Johnson, N. K., 1972 Origin and differentiation of the Avifauna of the Channel Islands, California. Condor 74: 295–315. https://doi.org/10.2307/1366591

Kratochwil, C. F., and A. Meyer, 2015 Closing the genotype-phenotype gap: emerging technologies for evolutionary genetics in ecological model vertebrate systems. BioEssays 37: 213–226. https://doi.org/10.1002/bies.201400142

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

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

Ma, Z., L. Li, C. Ye, M. Peng, and Y.-P. Zhang, 2018 Hybrid assembly of ultra-long Nanopore reads augmented with 10x-Genomics contigs: Demonstrated with a human genome. Genomics 111: 1896–1901.

Mason, N. A., P. O. Title, C. Cicero, K. J. Burns, and R. C. K. Bowie, 2014 Genetic variation among western populations of the Horned Lark (*Eremophila alpestris*) indicates recent colonization of the Channel Islands off southern California, mainland-bound dispersal, and postglacial range shifts. Auk 131: 162–174. https://doi.org/10.1642/AUK-13-181.1

Mason, N. A., and P. Unitt, 2018 Rapid phenotypic change in a native bird population following conversion of the Colorado Desert to agriculture. J. Avian Biol. 49: jav-01507. https://doi.org/10.1111/jav.01507

Shen, W., S. Le, Y. Li, and F. Hu, 2016 SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS One 11: e0163962.

Simão, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov, 2015 BUSCO: assessing genome assembly and annotation

completeness with single-copy orthologs. Bioinformatics 31: 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Trost, C. H., 1972 Adaptations of Horned Larks (*Eremophila alpestris*) to Hot Environments. Auk 89: 506–527.

Volker, M., N. Backstrom, B. M. Skinner, E. J. Langley, S. K. Bunzey *et al.*, 2010 Copy number variation, chromosome rearrangement, and their association with recombination during avian evolution. Genome Res. 20: 503–511. https://doi.org/10.1101/gr.103663.109

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.*, 2018 BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol. Biol. Evol. 35: 543–548. https://doi.org/10.1093/molbev/msx319

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

Communicating editor: C. Marshall