Genome Report: *De novo* assembly of a high-quality reference genome for the Horned Lark (*Eremophila alpestris*)

Nicholas A. Mason[1,2,3,*], Paulo Pulgarin[4,5], Carlos Daniel Cadena[4], Irby J. Lovette[1,2]

1. Fuller Evolutionary Biology Program, Cornell Lab of Ornithology, Cornell University, Ithaca, New York 14850
2. Department of Ecology and Evolutionary Biology, Cornell University, Ithaca, New York 14853
3. Museum of Vertebrate Zoology, University of California, Berkeley, California, 94720
4. Laboratorio de Biología Evolutiva de Vertebrados, Departamento de Ciencias Biológicas, Universidad de Los Andes, Bogotá, Colombia
5. Facultad de Ciencias y Biotecnología, Universidad CES, Medellín, Colombia

* Corresponding author: nmason@berkeley.edu

Keywords:

Alaudidae, ALLPATHS-LG, *Eremophila alpestris*, genome assembly, horned lark

Abstract

The Horned Lark (*Eremophila alpestris*) is a species of small songbird that exhibits remarkable geographic variation in appearance and habitat across an expansive distribution. While *E. alpestris* and related species have been the focus of many ecological and evolutionary studies, we still lack a highly contiguous genome assembly for horned larks and related taxa (Alaudidae). Here, we present CLO_EAlp_1.0, a highly contiguous assembly for horned larks generated from blood samples 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 2708 contigs with an N50 of 10.58 Mb and a L50 of 29. After polishing the genome, we were able to identify 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, indicating that our *de novo* assembly is near complete. 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 a widespread, yet understudied lineage of songbirds.

Introduction

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 horned larks 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 and 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; 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.

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 Tocá on the shores of the Embalse de La Copa in the Altiplano Cundiboyacense of the Boyacá department (5.623299˚ N, 73.184156˚ W). 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 underwent a population bottleneck upon colonizing the high-elevation plateaus of the region and therefore has high homozygosity, 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, 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 sizze using SeqKit v0.7.2 (Shen *et al.* 2016). We assessed the contiguity the *de novo* genome using QUAST v5.0.2 (Mikheenko *et al.* 2018) 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.

Mitochondrial Genome Assembly

We also assembled the mitochondrial genome for the same individual (EALPPER07) with NOVOplasty v3.7 (Hahn *et al.* 2013) using a ND2 sequence (GenBank Accession

KF743558) from a previous study (Mason *et al.* 2014) as the initial seed to begin the assembly process.

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 QUAST analyses are available from FigShare (doi:10.6084/m9.figshare.9956063;doi:10.6084/m9.figshare.9956042).

Results and Discussion

Taken together, the three lanes of Illumina HiSeq 2500 sequencing generated 1.59 x 10^9 total reads (~134x estimated coverage of a 1.2 Gb genome), including 5.45 x 10^8 paired-end reads for the 180 bp short-insert libraries, 1.24 x 10^8 paired-end reads for the 3 kb mate-pair library, and 1.27 x 10^8 paired-end reads for the 8 kb mate-pair library. Following scaffold polishing, the finalized CLO_EAlp_1.0 assembly consisted of 2708 contigs that totaled 1.04 Gb. The largest contig was 31.81 Mb while the N50 was 10.58 Mb and L50 was 29 (Table 1). The average GC content 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).

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). Thus, there is room to improve scaffolds generated in this assembly so that they match full chromosomes through strategies such as Hi-C (Burton *et al.* 2013) or ultra-long read sequencing technology (Ma *et al.* 2018). Functional annotation could also be improved by generating RNA-Seq and protein libraries for larks (Denoeud *et al.* 2008). Thus, while there is room to improve this current assembly, 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.

Acknowledgements

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

Author Contributions

NAM, PP, and CDC conducted field work. NAM performed lab work, genomic assembly, and analyses. NAM wrote the paper with input from PP, CDC, and IJL. All authors read and approved the final manuscript. The authors declare that they have no competing interests. The funding sponsors had no role in the design of the study; in the collection,

150 analyses, or interpretation of data; in the manuscript writing, and in the publication

151 decision.

152
153

References

Beason, R. C., 1995 Horned Lark (Eremophila alpestris), version 2.0. In The Birds of North America (A. F. Poole and F. B. Gill, Editors).

Behle, W., 1942 Distribution and variation of the Horned Larks (*Eremophila alpestris*) of western North America. University of California Publications in Zoology 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 versus traits paradox in birds. Molecular Biology and Evolution 34: 3123–3131.

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. Nature Biotechnolology 31: 1119–1125.

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 Research 18: 810–820.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos et al., 2009 BLAST+: architecture and applications. BMC Bioinformatics 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. Conservation Genetic Resources 7: 19–21.

Denoeud, F., J.-M. Aury, C. Da Silva, B. Noel, O. Rogier et al., 2008 Annotating genomes with massive-scale RNA sequencing. Genome Biology 9: R175.

175 Derjusheva, S., A. Kurganova, F. Habermann, and E. Gaginskaya, 2004 High
176 chromosome conservation detected by comparative chromosome painting in
177 chicken, pigeon and passerine birds. Chromosome Research 12: 715–723.
178 Dierickx, E., S. Sin, P. van Veelen, M. de L. Brooke, Y. Liu et al., 2019 Neo-sex
179 chromosomes and demography shape genetic diversity in the Critically
180 Endangered Raso lark: Genomics preprint https://doi.org/10.1101/617563
181 Donald, P. F., P. Alström, and D. Engelbrecht, 2017 Possible mechanisms of substrate
182 colour-matching in larks (Alaudidae) and their taxonomic implications. Ibis 159:
183 699–702.
184 Drovetski, S. V., S. F. Pearson, and S. Rohwer, 2006 Streaked horned lark *Eremophila*
185 *alpestris strigata* has distinct mitochondrial DNA. Conservervation Genetics 6:
186 875–883.
187 Drovetski, S. V., M. Raković, G. Semenov, I. V. Fadeev, and Y. A. Red'kin, 2014
188 Limited phylogeographic signal in sex-linked and autosomal loci despite
189 geographically, ecologically, and phenotypically concordant structure of mtDNA
190 variation in the Holarctic avian genus *Eremophila* (D. Mishmar, Ed.). PLoS ONE
191 9: e87570.
192 Ellegren, H., 2014 Genome sequencing and population genomics in non-model
193 organisms. Trends in Ecology & Evolution 29: 51–63.
194 Erickson, W. P., M. M. Wolfe, K. J. Bay, D. H. Johnson, and J. L. Gehring, 2014 A
195 comprehensive analysis of small-passerine fatalities from collision with turbines
196 at wind energy facilities (R. M. Brigham, Ed.). PLoS ONE 9: e107491.

Finn, R. D., J. Clements, and S. R. Eddy, 2011 HMMER web server: interactive sequence similarity searching. Nucleic Acids Research 39: W29–W37.

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. In press.

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. Proceedings of the National Academy of Sciences 108: 1513–1518.

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 Research 41: e129–e129.

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. The American Naturalist 188: 379–397.

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.

Johnson, N. K., 1972 Origin and differentiation of the avifauna of the Channel Islands, California. The Condor 74: 295–315.

Kratochwil, C. F., and A. Meyer, 2015 Closing the genotype-phenotype gap: emerging technologies for evolutionary genetics in ecological model vertebrate systems: Prospects & Overviews. BioEssays 37: 213–226.

Li, H., and R. Durbin, 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan et al., 2009 The sequence alignment/map format and SAMtools. Bioinformatics 25: 2078–2079.

Ma, Z. (Sam), L. Li, C. Ye, M. Peng, and Y.-P. Zhang, 2018 Hybrid assembly of ultra-long Nanopore reads augmented with 10x-Genomics contigs: D\demonstrated with a human genome. Genomics S0888754318305603.

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. The Auk 131: 162–174.

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.

Mikheenko, A., A. Prjibelski, V. Saveliev, D. Antipov, and A. Gurevich, 2018 Versatile genome assembly evaluation with QUAST-LG. Bioinformatics 34: i142–i150.

Shen, W., S. Le, Y. Li, and F. Hu, 2016 SeqKit: A Cross-platform and ultrafast toolkit for FASTA/Q File Manipulation (Q. Zou, Ed.). 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.

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

243 Volker, M., N. Backstrom, B. M. Skinner, E. J. Langley, S. K. Bunzey et al., 2010 Copy
244 number variation, chromosome rearrangement, and their association with
245 recombination during avian evolution. Genome Research 20: 503–511.
246 Walker, B. J., T. Abeel, T. Shea, M. Priest, A. Abouelliel et al., 2014 Pilon: An Integrated
247 Tool for Comprehensive Microbial Variant Detection and Genome Assembly
248 Improvement (J. Wang, Ed.). PLoS ONE 9: e112963.
249 Waterhouse, R. M., M. Seppey, F. A. Simão, M. Manni, P. Ioannidis et al., 2018 BUSCO
250 applications from quality assessments to gene prediction and phylogenomics.
251 Molecular Biology and Evolution 35: 543–548.
252 Wellenreuther, M., and L. Bernatchez, 2018 Eco-evolutionary genomics of
253 chromosomal inversions. Trends in Ecology & Evolution 33: 427–440.
254 de Zwaan, D. R., S. Barnes, and K. Martin, 2019 Plumage melanism is linked to male
255 quality, female parental investment and assortative mating in an alpine songbird.
256 Animal Behaviour 156: 41–49.
257

Table 1: De novo genome assembly metrics estimated using QUAST.

Assembly Statistic	CLO_EAlp_1.0
# contigs	2708
Largest contig	31807647
Total length	1041026391
GC (%)	42.23
N50 (bp)	10588015
N75 (bp)	3940266
L50	29
L75	70
# N's per 100 kbp	3472.39

262 Table 2: Output from BUSCO analyses to assess genome completeness by searching

263 for single-copy orthologs from aves and vertebrata datasets.

264

	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

265