DATA NOTE

Draft genome of tule elk *Cervus elaphus nannodes* [version 1; referees: awaiting peer review]

Jessica E. Mizzi\(^1\), Zachary T. Lounsberry\(^2\), C. Titus Brown\(^3\), Benjamin N. Sacks\(^4\)

\(^1\)Microbiology Graduate Group, University of California, Davis, CA, 95616, USA
\(^2\)Veterinary Genetics Laboratory, University of California, Davis, CA, 95616, USA
\(^3\)Department of Population Health and Reproduction, School of Veterinary Medicine, University of California, Davis, CA, 95616, USA
\(^4\)Department of Population Health and Reproduction and Mammalian Ecology and Conservation Unit, Veterinary Genetics Laboratory, School of Veterinary Medicine, University of California, Davis, CA, 95616, USA

Abstract

This paper presents the first draft genome of the tule elk (*Cervus elaphus nannodes*), a subspecies native to California that underwent an extreme genetic bottleneck in the late 1800s. The genome was generated from Illumina HiSeq 3000 whole genome sequencing of four individuals, resulting in the assembly of 2.395 billion base pairs (Gbp) over 602,862 contigs over 500 bp and N50 = 6,885 bp. This genome provides a resource to facilitate future genomic research on elk and other cervids.

Corresponding author: Benjamin N. Sacks (bnsacks@ucdavis.edu)

Author roles: Mizzi JE: Data Curation, Formal Analysis, Software, Writing – Original Draft Preparation, Writing – Review & Editing; Lounsberry ZT: Resources, Writing – Review & Editing; Brown CT: Resources, Supervision, Writing – Review & Editing; Sacks BN: Conceptualization, Funding Acquisition, Resources, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

How to cite this article: Mizzi JE, Lounsberry ZT, Brown CT and Sacks BN. Draft genome of tule elk *Cervus elaphus nannodes* [version 1; referees: awaiting peer review] *F1000Research* 2017, 6:1691 (doi: 10.12688/f1000research.12636.1)

Copyright: © 2017 Mizzi JE et al. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Grant information: Support for this project was provided by a grant to BNS from the California Department of Fish and Wildlife, FY1516 Big Game Management Program (Grant ID P1580009). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Introduction

To date, the closest genomic resource for elk (Cervus elaphus) is a full mitochondrial assembly of white-tailed deer (Odocoileus virginianus), a distantly related cervid. The present paper presents the first *de novo* genomic draft of the tule elk (C. elaphus nannodes). This California-endemic elk subspecies underwent a major genetic bottleneck when its numbers were reduced to as few as three individuals in the 1870s. Although their numbers have increased to >5,000 today, the historical bottleneck nevertheless left its mark on the elk’s genome, rendering it more homozygous than other elk subspecies.

Our motivation for generating a genomic resource for the tule elk was to create a reference for identifying single nucleotide polymorphisms (SNPs) to develop assays to monitor elk population abundance and for related population genetic applications. Due to the relatively low coverage generated in this work (40X overall with an average of 10X coverage from each individual), we used the MEGAHIT metagenome assembler, which has been found to perform well on low-quality or low-coverage DNA sequencing in bacteria.

Methods

Sample collection and library prep

Elk were selected from four geographically distinct populations across northern California to maximize genomic diversity (San Luis Reservoir, California Valley, American Canyon, and the San Luis National Wildlife Refuge). Genomic DNA was extracted from skin biopsies, which were obtained by the California Department of Fish and Wildlife as part of their elk management activities. We extracted DNA from skin using Qiagen DNeasy blood & tissue kits (QIAGEN Inc., Valencia, CA), according to the manufacturer’s instructions. The DNA was then fragmented via sonication using a Bioruptor (Diagenode, Denville, NJ) to 300 to 400 base pairs (bp) prior to adapter ligation. After verification of fragment size range using agarose gel electrophoresis, NEBNext® Ultra™ DNA Library Prep Kit for Illumina® (New England Biolabs, Inc., Ipswich, MA) was used to ligate Illumina adapters. Multiplexed libraries were prepared using NEBNext Multiplex Oligos for Illumina (New England Biolabs) to individually barcode each of four individual elk. Barcodes were annealed using low-cycle polymerase chain reactions during library preparation. To assess library quality, trace analysis was performed using a Bioanalyzer 2100 (Agilent, Santa Clara, CA) and fluorometric DNA quantitation of libraries was performed using a Qubit fluorometer (Invitrogen, Carlsbad, CA) prior to equilibrating sample concentrations and pooling for sequencing. After library quality control, four samples (one from each population) were pooled in equimolar concentrations and submitted for paired-end sequencing. Samples were sequenced on an Illumina HiSeq 3000 at the DNA Technologies and Expression Analysis Core of the UC Davis Genome Center.

Bioinformatics processing

Sequencing quality on demultiplexed reads was evaluated using FastQC v0.11.3 (RRID:SCR_014583). The Illumina TruSeq3-PE sequencing adapters were removed using Trimmomatic v0.30 (RRID:SCR_011848) with the ILLUMINACLIP parameter set to TruSeq3-PE.fa:2:40:15. The TruSeq3-PE.fa sequence was downloaded from https://anonscm.debian.org/cgit/debian-med/trimmomatic.git/plain/adapters/TruSeq3-PE.fa.

LEADING, TRAILING, and SLIDING parameters were set to 2, resulting in the removal of bases with a quality score of 2 or less according to a phred33 quality scoring matrix. The SLIDINGWINDOW parameter of 4:2 was used to clip reads once the quality score fell below 2 within the window. The MINLENGTH parameter set to 25 dropped any reads that fell below that length due to quality trimming. The demultiplexed, quality-filtered reads were interleaved using the interleave-reads.py script in khmer v2.0 (RRID:SCR_001156). The assembly was performed using MEGAHIT v1.0.5 on interleaved quality filtered reads. Genome statistical analysis was done using QUAST v3.0 (RRID:SCR_001228). All code used is publicly available at https://github.com/dib-lab/2017-tule-elk.

Results

We obtained 377,980,276 demultiplexed 150 bp paired-end raw reads, containing a total of 113.394 Gbp of sequence, or approximately 40X coverage of the approximately 3 Gbp tule elk genome. Sequence assembly resulted in the generation of a total genome sequence size of 2.395 Gbp. Reads were assembled into 602,862 contiguous sequences (“contigs”) averaging 3,973 bp in length with a minimum contig length of 201 bp. The G+C content of the genome was 41.55%. The N50 was 6,885 bp and maximum contig length was 72,391 bp. Additional assembly statistics are available in Table 1. No contigs (e.g. under a certain size or likely to reflect repeats) were removed from the assembly.

Table 1. Quality metrics on tule elk (Cervus elaphus nannodes) assembly, as generated with QUAST v3.0.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Tule elk assembly</th>
</tr>
</thead>
<tbody>
<tr>
<td># contigs ≥ 200 bp</td>
<td>1,367,218</td>
</tr>
<tr>
<td># contigs ≥ 500 bp</td>
<td>602,862</td>
</tr>
<tr>
<td># contigs ≥ 1000 bp</td>
<td>460,702</td>
</tr>
<tr>
<td># contigs ≥ 5000 bp</td>
<td>160,229</td>
</tr>
<tr>
<td># contigs ≥ 10000 bp</td>
<td>51,790</td>
</tr>
<tr>
<td># contigs ≥ 25000 bp</td>
<td>2,606</td>
</tr>
<tr>
<td># contigs ≥ 50000 bp</td>
<td>36</td>
</tr>
<tr>
<td>Total length ≥ 200 bp</td>
<td>2,607,088,486</td>
</tr>
<tr>
<td>Total length ≥ 1000 bp</td>
<td>2,295,163,580</td>
</tr>
<tr>
<td>Total length ≥ 5000 bp</td>
<td>1,531,314,985</td>
</tr>
<tr>
<td>Total length ≥ 10000 bp</td>
<td>771,863,493</td>
</tr>
<tr>
<td>Total length ≥ 25000 bp</td>
<td>80,157,993</td>
</tr>
<tr>
<td>Total length ≥ 50000 bp</td>
<td>2,056,962</td>
</tr>
<tr>
<td>Largest contig</td>
<td>72,391</td>
</tr>
<tr>
<td>Total length</td>
<td>2,395,105,945</td>
</tr>
<tr>
<td>GC</td>
<td>41.55%</td>
</tr>
<tr>
<td>N50</td>
<td>6,885</td>
</tr>
<tr>
<td>N75</td>
<td>3,646</td>
</tr>
<tr>
<td>L50</td>
<td>103,346</td>
</tr>
<tr>
<td>L75</td>
<td>222,107</td>
</tr>
<tr>
<td># N's per 100 kbp</td>
<td>0</td>
</tr>
</tbody>
</table>
This genome can serve as the basis for further genomic work on tule elk and other cervids, such as the development of a SNP assay to track elk population movement across increasingly developed northern Californian terrain. Furthermore, it is the first whole genome assembly available from the family Cervidae, providing a useful interim reference genome for bioinformatic analyses on other deer and elk species.

Data availability
Raw reads are available in the SRA under the BioProject ID PRJNA345218. The genome draft is available at https://doi.org/10.6084/m9.figshare.5382565.v111.

Code used in this study have been archived at http://doi.org/10.5281/zenodo.88793512

Competing interests
No competing interests were disclosed.

Grant information
Support for this project was provided by a grant to BNS from the California Department of Fish and Wildlife, FY1516 Big Game Management Program (Grant ID P150009).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements
JM would like to thank Luiz Irber, Camille Scott, and Lisa Johnson of the DIB lab at UC Davis for assistance with bioinformatics processing. We also thank C. Langner and J. Hobbs of the California Department of Fish and Wildlife for providing samples.

References