Title
Draft genome of tule elk Cervus elaphus nannodes

Authors
Mizzi, Jessica E
Lounsberry, Zachary T
Brown, C Titus
et al.

Publication Date
2017

DOI
10.12688/f1000research.12636.1

Peer reviewed

DATA NOTE

REVISED Draft genome of tule elk *Cervus canadensis nannodes* [version 2; referees: 2 approved]

Previously titled: Draft genome of tule elk *Cervus elaphus nannodes*

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

v2 **First published:** 15 Sep 2017, **6**:1691 (doi: 10.12688/f1000research.12636.1)
Latest published: 11 Dec 2017, **6**:1691 (doi: 10.12688/f1000research.12636.2)

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.

Open Peer Review

Referee Status: ✔ ✔

	Invited Referees 1	2
REVISED version 2 published 11 Dec 2017		✔ report
version 1 published 15 Sep 2017	✔ report	? report

1 **Steve Olsen**, National Animal Disease Center, ARS-USDA, USA

2 **Rudiger Brauning**, AgResearch, New Zealand

Discuss this article

Comments (0)

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 canadensis nannodes* [version 2; referees: 2 approved]** *F1000Research* 2017, **6**:1691 (doi: 10.12688/f1000research.12636.2)

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.

First published: 15 Sep 2017, **6**:1691 (doi: 10.12688/f1000research.12636.1)

REVISED **Amendments from Version 1**

In this version, the name *Cervus elaphus nannodes* was changed to *Cervus canadensis nannoodes* everywhere it appeared in the publication because most people now refer to the elk as *Cervus canadensis* to differentiate it from Eurasian red deer. Our original publication stated that we were presenting the first Cervidae genome, but this statement has been edited to reflect the recent addition (since our initial submission) of a red deer genome *Cervus elaphus hippelaphus* available on NCBI. Reference 1 has also been updated to point to this genome. The reported code in the "Bioinformatics processing" section contained an erroneous "SLIDING" parameter for trimmomatic, and this has been deleted to match the correct code on GitHub. Additional information about the quality of the sequencing run was added to the Results. Table 1 was reformatted for easier viewing.

See referee reports

Introduction

At the initiation of this project, no genome assembly existed for any member of the deer family (Cerivdae). We therefore sought to generate the first such assembly for the tule elk (*Cervus canadensis nannodes*). We note that after we completed our project and submitted the intial draft of this manuscript, a full assembly of red deer (*Cervus elaphus hippelaphus*) became available online[1]. The present paper presents the first *de novo* genomic draft of the tule elk. This California-endemic elk subspecies underwent a major genetic bottleneck when its numbers were reduced to as few as three individuals in the 1870s[2,3]. Although their numbers have increased to >5,000 today[4], 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[5].

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[4]). 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[4]. 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, the four samples 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)[6]. The Illumina TruSeq3-PE sequencing adapters were removed using Trimmomatic v0.30 (RRID:SCR_011848)[7] 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 and TRAILING 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)[8]. The assembly was performed using MEGAHIT v1.0.5[9] on interleaved quality filtered reads. Genome statistical analysis was done using QUAST v3.0 (RRID:SCR_001228)[10]. 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, from which 99.830 Gbp (88%) had quality scores ≥ Q30 (average quality score = 37.2), 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 canadensis nannodes*) assembly, as generated with QUAST v3.0.

Metric	Tule elk assembly
# contigs (≥ 200 bp)	1,367,218
# contigs ≥ 500 bp	602,862
# contigs (≥ 1000 bp)	460,702
# contigs (≥ 5000 bp)	160,229
# contigs (≥ 10000 bp)	51,790
# contigs (≥ 25000 bp)	2,606
# contigs (≥ 50000 bp)	36
Total length (≥ 200 bp)	2,607,088,486
Total length (≥ 1000 bp)	2,295,163,580
Total length (≥ 5000 bp)	1,531,314,985
Total length (≥ 10000 bp)	771,863,493
Total length (≥ 25000 bp)	80,157,993
Total length (≥ 50000 bp)	2,056,962
Largest contig	72,391
Total length	2,395,105,945
GC	41.55%
N50	6,885
N75	3,646
L50	103,346
L75	222,107
# N's per 100 kbp	0

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 one of the first whole genome assemblies 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.v1[11].

Code used in this study have been archived at http://doi.org/10.5281/zenodo.887935[12]

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 P1580009).

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

1. https://www.ncbi.nlm.nih.gov/genome/10790, accessed 11/29/17.
2. McCullough DR: **The Tule Elk: Its History, Behavior, and Ecology.** elibrary.ru. 1969.
Reference Source
3. Sacks BN, Lounsberry ZT, Kalani T, *et al.*: **Development and Characterization of 15 Polymorphic Dinucleotide Microsatellite Markers for Tule Elk Using HiSeq3000.** *J Hered.* 2016; **107**(7): 666–669.
PubMed Abstract | Publisher Full Text
4. Hobbs JH: **Draft Conservation and Management Plan for Elk. Report to the California Department of Fish and Wildlife.** 2014; 41.
5. Seitz A, Nieselt K: **Improving ancient DNA genome assembly.** *PeerJ.* 2017; **5**: e3126.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Andrews S: **FastQC: a quality control tool for high throughput sequence data.** 2010.
Reference Source
7. Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina sequence data.** *Bioinformatics.* 2014; **30**(15): 2114–20.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Crusoe MR, Alameldin HF, Awad S, *et al.*: **The khmer software package: enabling efficient nucleotide sequence analysis [version 1; referees: 2 approved, 1 approved with reservations].** *F1000Res.* 2015; **4**: 900.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Li D, Luo R, Liu CM, *et al.*: **MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices.** *Methods.* 2016; **102**: 3–11.
PubMed Abstract | Publisher Full Text
10. Gurevich A, Saveliev V, Vyahhi N, *et al.*: **QUAST: quality assessment tool for genome assemblies.** *Bioinformatics.* 2013; **29**(8): 1072–75.
PubMed Abstract | Publisher Full Text | Free Full Text
11. Mizzi J, Lounsberry ZT, Brown CT, *et al.*: **Tule Elk Draft Genome.** *figshare.* 2017.
Data Source
12. Mizzi J: **dib-lab/2017-tule-elk: Ready for publication.** *Zenodo.* 2017.
Data Source

Open Peer Review

Current Referee Status: ✔ ✔

Version 2

Referee Report 18 December 2017

doi:10.5256/f1000research.14577.r28939

✔ **Rudiger Brauning**
Invermay Agricultural Centre, AgResearch, Mosgiel, New Zealand

I'm happy with the changes made, no further comments.

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Referee Report 16 November 2017

doi:10.5256/f1000research.13682.r27606

? **Rudiger Brauning**
Invermay Agricultural Centre, AgResearch, Mosgiel, New Zealand

The authors describe the generation of a draft assembly for tule elk in the style of a brief genome announcement. For SNP detection and primer design this assembly is fine. It could e.g. be used in combination with Genotyping by Sequencing on additional individuals.

Materials and methods are sound and provided in full.

However a quick search of NCBI's taxonomy resource reveals that since June 2017 there is a genome assembly for red deer available https://www.ncbi.nlm.nih.gov/genome/10790. The authors therefore cannot claim to present the first whole genome assembly from the family Cervidae. Please change that statement.

Suggested further improvements:

Results

I would have liked to see a figure for the total amount of sequence after filtering as a simple way of

showing how good or bad the sequence run was.

Table 1's readability would be improved by getting all figures to align right.

I'd also recommend to add another assembly metric to look at the gene content; either using something like BUSCO or by mapping the refseq sequences of a related, well annotated species (e.g. cattle) against the draft genome.

Methods

Sample collection and library prep
I see that each individual has two tissue samples. The authors entered a sample ID into the 'tissue' field of NCBI's BioSample database. I'd recommend removing this and adding the animal ID in the 'isolate' field.

Please expand the entries in the 'isolation source' field. It says e.g. "Am. Cyn" which probably means American Canyon.

Bioinformatics processing
Checking the code I believe the statement "LEADING, TRAILING, and SLIDING parameters were set to 2" should read "LEADING and TRAILING parameters were set to 2".

Is the rationale for creating the dataset(s) clearly described?
Yes

Are the protocols appropriate and is the work technically sound?
Yes

Are sufficient details of methods and materials provided to allow replication by others?
Yes

Are the datasets clearly presented in a useable and accessible format?
Yes

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 05 Dec 2017

Jessica Mizzi, UC Davis, USA

Thank you for your review of this paper. Version 2 has been edited to reflect the presence of the red deer genome and a citation to that genome has been made. Table 1 has been reformatted for readability. The changes you've requested to the NCBI BioSample entry have been made. The trimmomatic code in the Bioinformatics Processing section has been edited to remove the erroneous "SLIDING" parameter. We've added text to the first sentence of the results section that describes the quality of sequence data in terms of standard quality scores. We opted not to provide

details on the gene content relative to a related genome as we felt this could be done more comprehensively in the future once the red deer genome has been published and peer-reviewed.

Competing Interests: I declare no competing interests.

Referee Report 25 October 2017

doi:10.5256/f1000research.13682.r26607

Steve Olsen
Infectious Bacterial Diseases Research Unit, National Animal Disease Center, ARS-USDA, Ames, IA, USA

This article describes the generation of a draft genome (40X coverage from 4 animals) of the tule elk (Cervus elaphus nannodes). The research methods are fairly standard for the Illumina sequencing used. At 602,862 contigs, the genome is very prelminary and will require quite a bit of additional work in order for it to be applicable to a wide range of applications. The report basically falls into a category of a genome announcement.

Is the rationale for creating the dataset(s) clearly described?
Yes

Are the protocols appropriate and is the work technically sound?
Yes

Are sufficient details of methods and materials provided to allow replication by others?
Yes

Are the datasets clearly presented in a useable and accessible format?
Yes

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 05 Dec 2017

Jessica Mizzi, UC Davis, USA

Thank you for your review of this paper.

Competing Interests: I declare no competing interests.