Data in Brief 41 (2022) 107857

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Draft genome sequence data of Indian rhinoceros, *Rhinoceros unicornis*

Kei Nabeshima[a], Nobuyoshi Nakajima[b], Mitsuaki Ogata[c], Manabu Onuma[a,*]

[a] *Ecological Risk Assessment and Control Section Center for Environmental Biology and Ecosystem, National Institute for Environmental Studies, 16-2 Onogawa, Tsukuba, Ibaraki 305-8506, Japan*
[b] *Environmental Genomics Office Center for Environmental Biology and Ecosystem, National Institute for Environmental Studies, 16-2 Onogawa, Tsukuba, Ibaraki 305-8506, Japan*
[c] *Preservation and Research Center of Yokohama City, 155−1 Kawaisyuku, Asahi−ku, Yokohama, Kanagawa 241−0804, Japan*

ARTICLE INFO

Article history:
Received 5 November 2021
Revised 13 January 2022
Accepted 18 January 2022
Available online 22 January 2022

Keywords:
Indian rhinoceros
Whole-genome sequence
Wildlife
Hybrid sequencing

ABSTRACT

The Indian rhinoceros (*Rhinoceros unicornis*) is a large herbivore found in northern India and southern Nepal. It is a critically endangered species, with an estimated population of approximately 3,600 in the wild. Genetic factors, such as the loss of genetic diversity and the accumulation of deleterious variations, are critical risk factors for the extinction of endangered species, such as the Indian rhinoceros. To support the conservation efforts of the Indian rhinoceros, we assembled its draft genome. The new genomic data will enable the study of functional genes associated with the ecological and physiological characteristics of Indian rhinoceros and help us establish more effective conservation measures. The muscles of an Indian rhinoceros that died from prostration at a zoo were collected, and the samples were stored at the National Institute for Environmental Studies (Tsukuba, Japan). Sequence data were obtained using an Illumina NovaSeq 6000 platform for short reads and an Oxford Nanopore Technologies PromethION for long reads. We generated approximately 235.2 Gbp of data. From these sequences, we assembled a 2,375,051,758 bp genome consisting of 7,615 contigs. The genome data are available from the National Center

* Corresponding author.
E-mail address: monuma@nies.go.jp (M. Onuma).

https://doi.org/10.1016/j.dib.2022.107857

Biotechnology Information BioProject database under accession number BOSQ00000000.

Specifications Table

Subject	Biodiversity
Specific subject area	Genomics
Type of data	Genome sequences and table
How the data were acquired	High-throughput DNA sequencing using NovaSeq 6000 and PromethION platforms
Data format	Raw and assembled genome sequences
Description of data collection	The sample was obtained from the muscle tissue of *Rhinoceros unicornis* at Yokohama Municipal Kanazawa Zoo, Yokohama, Japan (NIES ID: 5488M, female). Genomic DNA was extracted using proteinase K and phenol/chloroform/isoamyl alcohol for short-read sequencing and a NucleoBond HMW DNA extraction kit (Macherey-Nagel, Düren, Germany) for long-read sequencing. Short-read libraries were prepared using a TruSeq LT PCR-free DNA Library Preparation Kit, and sequencing was performed using the NovaSeq 6000 sequencing system (Illumina, San Diego, CA, USA) with 2 × 150 bp paired-end reads. Long-read libraries were prepared using a Ligation Sequencing Kit, and sequencing was performed using the PromethION system (Oxford Nanopore Technologies, Oxford, UK). The short and long reads were assembled into contigs using the HASLR program, which utilizes a hybrid assembly approach.
Data source location	Tsukuba, Ibaraki, Japan
Data accessibility	Data have been deposited in relevant databases and are publicly available. The sequencing data were deposited in the Sequence Read Archive under accession numbers DRR308100 (https://www.ncbi.nlm.nih.gov/sra/?term=DRR308100) and DRR311486 (https://www.ncbi.nlm.nih.gov/sra/?term=DRR311486). The whole-genome sequence, *Rhinoceros unicornis* ID: 5488M, was deposited in GenBank under accession number BOSQ00000000 (https://www.ncbi.nlm.nih.gov/nuccore/2085786713). All details regarding genome sequencing data are available at NCBI under BioProject accession number PRJDB11285 (https://www.ncbi.nlm.nih.gov/bioproject/PRJDB11285).

Value of the Data

- The Indian rhinoceros (*Rhinoceros unicornis*) is a critically endangered herbivore with little genetic information available.
- The Indian rhinoceros genome data can be used for *ex situ* conservation and infectious disease control.
- These genome data can be analyzed together with other high-quality Indian rhinoceros whole-genome data (accession number: JAFHKO000000000) to assess the diversity of the Indian rhinoceros species and identify genomic rearrangements.
- These data may also be used by other researchers to identify genes related to immunity and plan breeding programs to maintain genetic diversity.

1. Data Description

The Indian rhinoceros (*Rhinoceros unicornis*) is an endangered species categorized as vulnerable by the International Union for Conservation of Nature Redlist of Threatened Species [1]. Although the species declined to near extinction in the early 1900s, the population of Indian

Table 1
Amount of data generated.

Type of reads	No. of reads	Average Read length	Total data
Short	623,580,225	150 bp	188.3 Gbp
Long	5,536,969	8,236.2 bp	46.9 Gbp

Table 2
General features of the *Rhinoceros unicornis* genome.

GC content (%)	40.99
Number of contigs	7,615
Number of scaffolds	7,615
Total contig length (bp)	2,375,051,758
N50 contig size (bp)	663,630
Longest sequence (bp)	5,292,610
Shortest sequence (bp)	10,012
Mean sequence length (bp)	311,891
Median sequence length (bp)	156,082
BUSCO score	C:96.5% [S:96.1%, D:0.4%], F:3.0%, M:0.5%, n:233

rhinoceros is currently increasing. The total population estimate in August 2018 was 3,588 individuals, with 649 animals in Nepal and 2,939 in India [2]. However, despite the increasing population size, there are still threats to the species [2].

It is important to consider genetic factors for conservation activities because the Indian rhinoceros declined to near extinction in the early 1900s [2]. Genetic factors such as the loss of genetic diversity and the accumulation of deleterious variations are known to be critical risk factors for the extinction of endangered species [3–6]. To support the conservation efforts of the Indian rhinoceros, we generated high-precision genomic data. These data will enable the study of functional genes associated with the ecological and physiological characteristics of the Indian rhinoceros and will facilitate the establishment of more effective conservation measures.

We sequenced both short-read and long-read libraries and generated approximately 235.2 Gbp of data. We obtained 623,580,225 short reads and 5,536,969 long reads (Table 1). The sequencing data were deposited in the Sequence Read Archive under accession numbers DRR308100 and DRR311486. The short and long reads were assembled into contigs using the HASLR program, which utilizes a hybrid assembly approach [7]. We assembled a 2,375,051,758 bp genome consisting of 7,615 contigs, with an N50 of 663 kbp. The GC content of the Indian rhinoceros was 40.99%, and the complete Benchmarking Universal Single-Copy Orthologs (BUSCO) score (C) was 96.5% (single copy, S:96.1%; duplicated, D:0.4%; fragmented, F:3.0%; and missing, M:0.5%; Table 2). The Indian rhinoceros genome sequence is a potentially useful resource for future molecular evolutionary analyses of mammals.

2. Experimental Design, Materials and Methods

2.1. Sample preparation and sequencing

We sampled *Rhinoceros unicornis* at Yokohama Municipal Kanazawa Zoo, Yokohama, Japan (NIES ID: 5488M, female). The rhinoceros was born on February 1, 2007 and died from prostration on March 22, 2007. Muscle tissue was autopsied to determine the cause of death. Genomic DNA was extracted from the muscles using proteinase K and phenol/chloroform/isoamyl alcohol for short-read sequencing and a NucleoBond HMW DNA extraction kit for long-read sequencing. Short-read whole-genome sequencing was performed by Macrogen Japan (Tokyo,Japan). Short-read libraries were prepared using a TruSeq LT PCR-free DNA Library Preparation Kit, and sequencing was performed using the NovaSeq 6000 sequencing system, with 2×150 bp

paired-end reads. Long-read sequencing was performed by GeneBay (Yokohoma, Japan). Long-read libraries were prepared using a Ligation Sequencing Kit, and sequencing was performed using the PromethION system.

2.2. De novo *genome assembly and assessment*

The sequenced reads were assembled using HASLR v. 2020-06a, which utilizes a hybrid assembly approach [7]. Assembly was performed by specifying a minimum long-read coverage of 20 ×, an estimated genome size of 2.4 Gb, and other parameters kept at default settings. The genome assembly was evaluated using BUSCO v5 [8], based on the core vertebrate gene set [9], using the gVolante pipeline [10,11].

Ethics Statement

None.

CRediT Author Statement

Kei Nabeshima: Methodology, Software, Writing– original draft preparation; **Nobuyoshi Nakajima:** Data curation; **Mitsuaki Ogata:** Sampling; **Manabu Onuma:** Conceptualization, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that have or could be perceived to have influenced the work reported in this article.

Acknowledgments

We would like to thank the staff at Macrogen Japan and GeneBay for their careful handling of our samples.

References

[1] IUCN, The IUCN Red List of Threatened Species. Version 2021-2. http://www.iucnredlist.org. Accessed October 6, 2021.
[2] S. Ellis, B. Talukdar, *Rhinoceros unicornis*. https://dx.doi.org/10.2305/IUCN.UK.2019-3.RLTS.T19496A18494149.en. Accessed October 6, 2021.
[3] P.A. Hohenlohe, W.C. Funk, O.P. Rajora, Population genomics for wildlife conservation and management, Mol. Ecol. 30 (2021) 62–82, doi:10.1111/mec.15720.
[4] M. Lynch, J. Conery, R. Burger, Mutation accumulation and the extinction of small populations, Am. Nat. 146 (1995) 489–518 http://www.jstor.org/stable/2462976.
[5] M.A. Supple, B. Shapiro, Conservation of biodiversity in the genomics era, Genome Biol. 19 (2018) 131, doi:10.1186/s13059-018-1520-3.
[6] J.C. Teixeira, C.D. Huber, The inflated significance of neutral genetic diversity in conservation genetics, Proc. Nat. Acad. Sci. U.S.A. 118 (2021) e2015096118, doi:10.1073/pnas.2015096118.
[7] E. Haghshenas, H. Asghari, J. Stoye, C. Chauve, F. Hach, HASLR: fast hybrid assembly of long reads, iScience 23 (2020) 101389, doi:10.1016/j.isci.2020.101389.
[8] M. Manni, M.R. Berkeley, M. Seppey, F.A. Simão, E.M. Zdobnov, BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes, Mol. Biol. Evol. 38 (2021) 4647–4654, doi:10.1093/molbev/msab199.

[9] Y. Hara, K. Tatsumi, M. Yoshida, E. Kajikawa, H. Kiyonari, S. Kuraku, Optimizing and benchmarking de novo transcriptome sequencing: from library preparation to assembly evaluation, BMC Genomics 16 (2015) 977, doi:10.1186/s12864-015-2007-1.
[10] O. Nishimura, Y. Hara, S. Kuraku, gVolante for standardizing completeness assessment of genome and transcriptome assemblies, Bioinformatics 33 (2017) 3635–3637, doi:10.1093/bioinformatics/btx445.
[11] O. Nishimura, Y. Hara, S. Kuraku, Evaluating genome assemblies and gene models using gVolante, in: M Kollmar (Ed.), Gene Prediction, Methods in Molecular Biology, vol 1962, Humana, New York, 2009, pp. 247–256, doi:10.1007/978-1-4939-9173-0_15.