DATA NOTE | Article in Press

Short read assembly and gene annotation of the first genome of the extinct Indian Javan subspecies, *Rhinoceros sondaicus inermis*

Sirithorn Kongseng · Putita Jiranuntskul · Yada Katanyuphan · Pipad Krajaejun · Athiwat Wattanapituksakul · Wunrada Surat

Received: 24 February 2026 / Accepted: 17 September 2026
© The Author(s) 2026

Abstract

Objective The Javan rhino (*Rhinoceros sondaicus*) is a critically endangered species in which approximately 50 individuals survive in the Ujung Kulon National Park in Southwest Java, Indonesia. There is limited genetic data of this rhino species in which only mitochondrial DNA sequences are available in the GenBank database. Here, the whole genome of the Javan rhino was sequenced, and the draft genome was assembled and annotated.

Data description The whole genome sequencing was performed using a tooth of the ancient Indian Javan rhino dated approximately 100 years before present collected from the rhino skeleton collection room in Khao Sam Roi Yot National Park, Prachuap Kiri Khan Province, Thailand. A total of 221 million reads were generated, and 51 million reads were mapped to black rhino (*Diceros bicornis*) sequences. Then, the total length of 195 Mb was assembled into 234,360 contigs and 343 scaffolds. The genome annotation was performed by comparing to the NCBI annotation of black rhino sequences which resulted in more than 6,000 annotated genes. This genomic data could gain insight into physiological characteristics, and it would be useful for evolutionary studies and conservation efforts of the critically endangered rhino species.

Keywords *Rhinoceros sondaicus* · Javan rhino · Whole genome sequencing · Genome assembly · Gene annotation · Ancient DNA · Thailand

Objective

Genomic data has played a significant role for animal conservation [1]. The study of genomic DNA could provide insight into the ecological and physiological characteristics of endangered animal species which may relate to climate adaptation, health issues and fertility. Hence, the whole genome sequencing of endangered species is very important for conservation efforts.

The Javan rhino (*Rhinoceros sondaicus*) is a critically endangered species which the main cause of its population decline is rhino poaching for horns [2]. In the past, the rhino species had distributed into several countries in Asia, but it had become extinct from Thailand decades ago [2]. Recently, the Javan rhinos are found only in

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.

Published online: 22 September 2026

Table 1 Overview of data files/ data sets

Label	Name of data file/ data set	File types (file extension)	Data repository and identifier (DOI or accession number)
Data file 1	Genome assembly of *R. sondaicus*	Fasta file (.fna)	DDBJ Genome Assembly, accession BAAKGJ010000001–BAAKGJ010000343 [8]
Data file 2	Gene annotation	Gff file (.gff)	Figshare, https://doi.org/10.6084/m9.figshare.29986387 [9]
Data set 1	Short read sequences of *R. sondaicus*	Fastq files (.fq.gz)	Sequence Read Archive (SRR35152370) https://identifiers.org/ncbi/insdc.sra:SRR35152370 [10]

the Ujung Kulon National Park in Southwest Java, Indonesia, in which approximately 50 individuals survive [2]. There are three Javan rhino subspecies; *R. sondaicus sondaicus* (the Indonesian Javan rhino), *R. sondaicus inermis* (the Indian Javan rhino) and *R. sondaicus annamiticus* (the Vietnamese Javan rhino), which only the first one is an extant subspecies [3, 4]. The Vietnamese and Indian Javan rhinos became extinct in 2010 and the early nineteenth century, respectively [5, 6]. Up to date, there is a few genetic data of the extinct Javan rhino subspecies in which only three mitochondrial DNA sequences are available in the NCBI database. The study of genomic DNA could provide insight into the ecological and physiological characteristics of the rhino species which could be useful for evolutionary studies and conservation efforts of this rhino species.

We present the results of whole genome sequencing, assembly and gene annotation of a tooth from the ancient Indian Javan rhino from Thailand dated approximately 100 years before present.

Data description

An ancient rhino specimen of the extinct Indian Javan rhino subspecies dated to approximately 100 years before present was collected from the rhino skeleton collection room in the tourist service centre, Khao Sam Roi Yot National Park, Prachuap Khiri Khan Province, Thailand [7]. The tooth was surface cleaned using a dental drill and exposed to UV, then it was ground into fine powder. Genomic DNA was extracted from the powder using EDTA/proteinase K and Qiagen MinElute PCR Purification Kit (Qiagen, Germany) following the manufacturer's instructions. The genomic library was prepared using the VAHTS Universal Plus DNA Library Prep Kit for Illumina V2 (Vazyme, China) and sequenced using the NovaSeq 6000 with 2 × 150 bp paired-end reads. This resulted in 221 million reads and the total data of 33 Gb (Table 1 dataset 1, [10]). Sequence quality control and adapter trimming was preceded using Trimmomatic v0.39 [11]. The cleaned read quality was inspected by FASTQC [12]. Then cleaned reads were aligned to the black rhino genome assembly; accession GCA_013634535.1 [13] and 51 million reads were mapped to the rhino genome using BWA [14]. Subsequently, damage patterns in ancient DNA sequences were analysed by mapDamage v2.0 [15]. The result showed very low substitutions (< 1%) of C to T at 5' and G to A at 3' ends which may be due to young age and very well-preserved condition of the ancient specimen. Then, the mapped reads were assembled into contigs and scaffolds using SPAdes v4.0.0 and RagTag v2.1.0 [16, 17]. This resulted in a draft genome assembled of approximately 195 Mb, containing 234,360 contigs with a contig N50 of 1,989 bp and 343 scaffolds (accession No. BAAKGJ010000001–BAAKGJ010000343) which the longest scaffold is 1,436,825 bp (Table 1, data file 1, [8]). Gene prediction and annotation of the draft genome was performed by comparing to the NCBI annotation of black rhino (version 8 Feb 2025) using AUGUSTUS [18] which resulted in more than 6,000 annotated genes (Table 1, data file 2, [9]).

Limitations

Here, we used the ancient specimen from a tropical region, Thailand, where high temperature and high humidity can destroy most of the DNA in it. The draft genome of the ancient rhino specimen has not been fully covered.

Abbreviations

BWA	Burrows-Wheeler Aligner
DNA	Deoxyribonucleic acid
EDTA	Ethylenediaminetetraacetic acid
FASTQC	FAST Quality Control
Gb	Giga bases
NCBI	National Center Biotechnology Information
PCR	Polymerase Chain Reaction
UV	Ultraviolet

Acknowledgements This project was funded by the National Research Council of Thailand (N42A650275) and Kasetsart University, Bangkok, Thailand (B4301-00121-65). We would like to thank the Department of National Parks, Wildlife and Plant Conservation and Mr. Pongsatorn Promkuntod, Chief of Khao Sam Roi Yot National Park for providing the rhino specimen. The publication fee was supported by International SciKU Branding (ISB), Faculty of Science, Kasetsart University.

Author contributions SK analysed genomic data, PJ analysed damage pattern by mapDamage, YK performed the DNA extraction, PK and AW prepared documents to get the specimen from the National park, WS coordinated the project, decided the work and wrote the manuscript. The authors read and approved the final manuscript.

Funding This project was funded by the National Research Council of Thailand (N42A650275) and Kasetsart University, Bangkok, Thailand (B4301-00121-65). We would like to thank the Department of National Parks, Wildlife and Plant Conservation and Mr. Pongsatorn Promkuntod, Chief of Khao Sam Roi Yot National Park for providing the rhino specimen. The publication fee was supported by International SciKU Branding (ISB), Faculty of Science, Kasetsart University.

Data availability The data described in this Data Note can be freely and openly accessed on Sequence Read Archive (SRR35152370) under Bioproject ID PRJNA1306765 (http://identifiers.org/bioproject:PRJNA1306765) and BioSample ID SAMN50645368 (http://identifiers.org/BioSample:SAMN50645368) [10], assembled sequences under accession number BAAKGJ010000001–BAAKGJ010000343 [8] and figshare (https://doi.org/10.6084/m9.figshare.29986387) [9]. Please see Table 1 for details and links to the data.

Declarations

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

References

1. van der Valk T, Dalén L. From genomic threat assessment to conservation action. Cell. 2024;187:1038–41. https://doi.org/10.1016/j.cell.2024.01.038.
2. International Rhino Foundation. https://rhinos.org/about-rhinos/state-of-the-rhino/. Accessed on 21 July 2025.

3. Wilson D, Reeder D. Mammal species of the world. Baltimore, MD, USA: Johns Hopkins University; 2005.
4. Groves CP, Leslie DM. *Rhinoceros sondaicus* (Perissodactyla: Rhinocerotidae). Mammalian Species. 2011;437:190–208.
5. Brook SM, van Coeverden de Groot P, Scott C, et al. Integrated and novel survey methods for rhinoceros populations confirm the extinction of *Rhinoceros sondaicus annamiticus* from Vietnam. Biol Conserv. 2012;155:59–67.
6. Rookmaaker LC. The distribution of the rhinoceros in eastern India, Bangladesh, China, and the Indochina region. Zool Anz. 1980;205:253–68.
7. Katanyuphan Y, Krajaejun P, Wattanapituksakul A, Surat W. The first report of rhino DNA in Thailand: A possible extinct Indian Javan subspecies, *Rhinoceros sondaicus inermis*. Animals. 2025;15:1678. https://doi.org/10.3390/ani15121678.
8. Assembled sequences. 2026. https://identifiers.org/ncbi/insdc:BAAKGJ010000001 to https://identifiers.org/ncbi/insdc:BAAKGJ010000343.
9. Kongseng S, Katanyuphan Y, Krajaejun P, Wattanapituksakul A, Surat, Wunrada S. Annotation of the first Javan rhino from Thailand. 2026. https://doi.org/10.6084/m9.figshare.29986387.v1.
10. Sequence Read Archive. 2026. https://identifiers.org/ncbi/insdc.sra:SRR35152370.
11. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20. https://doi.org/10.1093/bioinformatics/btu170.
12. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
13. Moodley Y, Westbury MV, Russo I-RM, Gopalakrishnan S, Rakotoarivelo A, Olsen R-A, et al. Interspecific Gene Flow and the evolution of specialization in Black and White Rhinoceros. Mol Biol Evol. 2020;37:3105–17.
14. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60. https://doi.org/10.1093/bioinformatics/btp698.
15. Jónsson H, Ginolhac A, Schubert M, Johnson P, Orlando L. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics. 2013;29(13):1682–4. https://doi.org/10.1093/bioinformatics/btt193.
16. Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. Using SPAdes de novo assembler. Curr Protoc Bioinf. 2020;70:e102. https://doi.org/10.1002/cpbi.102.
17. Alonge M, Lebeigle L, Kirsche M, Jenike K, Ou S, Aganezov S, et al. RagTag: A software package for improving incomplete genome assemblies with whole-genome alignments. PLOS Comput Biol. 2022;18:e1009730. https://doi.org/10.1371/journal.pcbi.1009730.
18. Stanke M, Waack S. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics. 2003;19:ii215–25. https://doi.org/10.1093/bioinformatics/btg1080.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Sirithorn Kongseng[1] · Putita Jiranuntskul[1] · Yada Katanyuphan[1] · Pipad Krajaejun[2] · Athiwat Wattanapituksakul[3] · Wunrada Surat[1]

✉ Wunrada Surat
fsciwrds@ku.ac.th

Sirithorn Kongseng
sirithorn.kon@ku.th

Putita Jiranuntskul
fscipij@ku.ac.th

Yada Katanyuphan
yadoze@gmail.com

Pipad Krajaejun
pipad_k@yahoo.com

Athiwat Wattanapituksakul
essathiwat@gmail.com

[1] Department of Genetics, Faculty of Science, Kasetsart University, Bangkok 10900, Thailand

[2] Department of History, Faculty of Liberal Arts, Thammasat University, Pathum Thani 12120, Thailand

[3] The Timing and the Cause of the Transition in Domestic Cattle Species in Thailand Project, Department of Genetics, Faculty of Science, Kasetsart University, Bangkok 10900, Thailand