scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A reference genome assembly of *Pteropus pselaphon*

Kei Nabeshima[1] ✉, Yasuhito Shimada[2,3] & Manabu Onuma[1]

The Bonin flying fox (*Pteropus pselaphon*) is an endemic species of the Ogasawara Islands, Japan, and is classified as Critically Endangered. Here, we report a genome assembly of *P. pselaphon*, constructed using Nanopore long reads and Omni-C scaffolding. The final assembly spans 2.16 Gb across 929 scaffolds and achieves a Benchmarking Universal Single-Copy Orthologs (BUSCO) completeness score of C:98.2% [S:94.3%, D:4.0%], F:0.7%, M:1.0%. Gene prediction with Helixer software identified 20,647 protein-coding sequences, with a BUSCO score of C:92.2% [S:88.9%, D:3.3%], F:3.3%, M:4.5%. This genome provides a valuable resource for the conservation management of this endangered species, as well as for evolutionary genomics and studies of host–pathogen interactions in flying fox.

Background & Summary

The Bonin flying fox (*Pteropus pselaphon*; Fig. 1) is an island-endemic bat restricted to the Ogasawara Islands of Japan and is currently listed as Critically Endangered (CR) on the IUCN Red List[1]. Historical records indicate substantial population declines during the 20th century associated with human pressures, and the species has since been the subject of strengthened legal protection and conservation actions in Japan[2]. It is threatened primarily by habitat loss and invasive species, and recent assessments estimate a remaining population of approximately 100–200 individuals[3]. To our knowledge, no genome-wide phylogenomic study has been reported for *P. pselaphon* to date; however, population genetic analyses based on mitochondrial DNA and a limited number of nuclear markers have been conducted[4]. The reference genome presented here provides a foundational resource for future conservation genomic and comparative evolutionary studies of this insular endemic. The Ogasawara Islands, located in the northwestern Pacific Ocean about 1,300 km from Okinawa, 1,400 km from the Mariana Islands, and 850 km from mainland Japan, are oceanic volcanic islands that emerged approximately 48 million years ago. It is generally believed that flying foxes can sustain continuous flight over distances of roughly 100 km at most, and it remains unclear how the Bonin flying fox reached the Ogasawara Islands. The timing of its arrival in the Ogasawara Islands remains uncertain, but genetic and biogeographic evidence suggests a sufficiently long history to have diverged from any of the populations in the Philippines, Okinawa, and the Mariana Islands. Bats of the genus *Pteropus* have served as important models for studies of immunity, virus–host coevolution, longevity, and metabolism. However, no high-quality genome resources have previously been available for *P. pselaphon*, restricting both conservation genomic approaches and broader comparative analyses. Island-endemic populations are particularly vulnerable to genetic erosion, underscoring the urgent need for a reference genome. Here we present an assembly for *P. pselaphon*, generated using Oxford Nanopore PromethION long-read sequencing (69 Gb; read N50 = 29,987 bp) in combination with Omni-C chromatin conformation capture (12.3 Gb). The final assembly spans 2.16 Gb across 929 scaffolds and achieves a BUSCO completeness score of C:98.2% [S:94.3%, D:4.0%], F:0.7%, M:1.0%. Gene prediction with Helixer identified 20,647 protein-coding sequences, with a BUSCO score of C:92.2% [S:88.9%, D:3.3%], F:3.3%, M:4.5%. This genome provides a valuable resource for the conservation management of this endangered species, as well as for evolutionary genomics and studies of host–pathogen interactions in bats.

Methods

Sample collection and ethical approval. The *Pteropus pselaphon* individual analyzed in this study was an adult male that was rescued on 8 August 2021 on Chichi-jima Island in the Ogasawara Islands, Japan. It was subsequently diagnosed with methicillin-resistant *Staphylococcus aureus* (MRSA) infection. Because release and long-term captivity were not feasible, the bat was humanely euthanized with approval from the Ministry of the

[1]Biodiversity Division, Biodiversity Resource Conservation Office, National Institute for Environmental Studies, 16-2 Onogawa, Tsukuba, Ibaraki, 305-8506, Japan. [2]Department of Integrative Pharmacology, Mie University Graduate School of Medicine, 2-174 Edobashi, Tsu, Mie, 514-8507, Japan. [3]Mie University Zebrafish Research Center, 2-174 Edobashi, Tsu, Mie, 514-8507, Japan. ✉e-mail: nabeshima.kei@nies.go.jp

Fig. 1 Photograph of the *Pteropus pselaphon*. A wild Bonin flying fox (*Pteropus pselaphon*) inhabiting the Ogasawara Islands, Japan (photograph by Seiji Narita, National Institute for Environmental Studies).

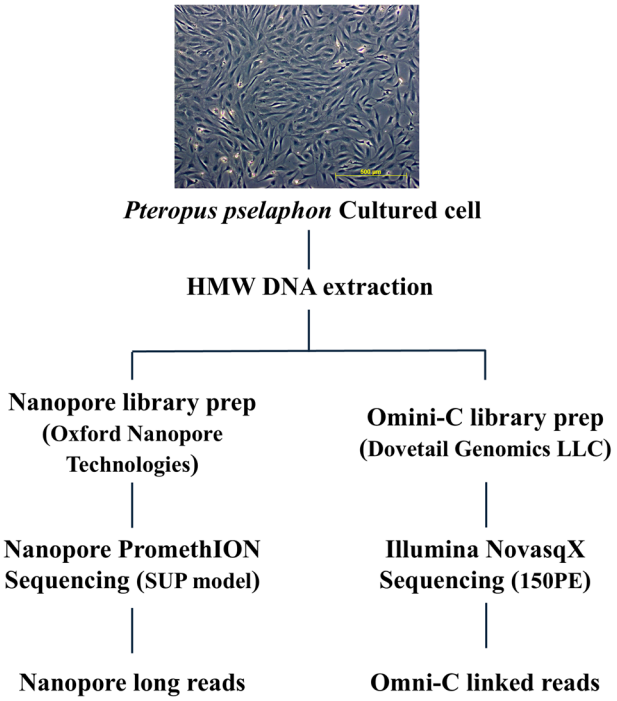

Fig. 2 Experimental workflow and generated sequencing datasets for *Pteropus pselaphon*.

Environment, Japan, on 19 May 2023. Wing membrane tissue was collected immediately after euthanasia for primary cell isolation and subsequent genomic DNA extraction. The primary cell culture is preserved in the NIES Time Capsule Project (cell line ID: 7782 M). All methods were carried out in accordance with relevant guidelines and regulations. This study is reported in accordance with ARRIVE guidelines (https://arriveguidelines.org).

DNA extraction and sequencing. The sequence data acquisition workflow is summarized in Fig. 2. For Nanopore sequencing, high-molecular-weight genomic DNA was isolated using a standard phenol–chloroform extraction protocol. DNA quality and integrity were assessed with NanoDrop spectrophotometry, Qubit fluorometry (Thermo Fisher Scientific, MA, USA), and pulsed-field gel electrophoresis. Ligation Sequencing Kit V14 (SQK-LSK114; Oxford Nanopore Technologies, Oxford, UK) libraries were constructed following the standard protocol of Nanopore PromethION using a single R10.4.1 flow cell (FLO-PRO114M). Raw signal data were basecalled with Dorado (v7.6) using the SUP model. Prior to assembly, Nanopore reads were filtered to retain only high-quality reads (quality score $\geq$ Q10). The quality of the Nanopore reads was evaluated using NanoPlot v1.42.0[5]. A total of approximately 5.3 million reads (about 69 Gb) were obtained, with a read length N50 of 29,987 bp, corresponding to approximately 35 $\times$ coverage assuming an estimated genome size of 2 Gb (Table 1).

For Omni-C sequencing, Omni-C chromatin conformation capture data were generated following the Dovetail Genomics (Dovetail Genomics LLC, CA, USA) protocol and sequenced on an Illumina NovaSeq X

	Nanopore reads	Omni-C reads
Total bases	69.882 Gb	12.255 Gb
Number of reads	5.317 M	81.706 M
Mean read length (bp)	13142.4	150 bp paired-end reads
Median read length (bp)	5127	
Reads quality	N50: 29,987 > Q10 reads: 99.6% > Q20 reads: 68.8%	>Q20 reads: 95.85%
		PCR duplication rate:28.84%
Mapping rate	99.6%	98.2%

Table 1. Read statistics.

	Genome	CDSs
Number	929	20,647
Size	2,167,613,222	35,068,977
Mean length	2333275.8	1698.50
N50/L50	81982708/11	2355/3986
BUSCO score (mammalia_odb10)	C:98.2% [S:94.3%, D:4.0%], F:0.7%, M:1.0%	C:92.2% [S:88.9%, D:3.3%], F:3.3%, M:4.5%

Table 2. Evaluation of the genome and coding sequences of *Pteropus pselaphon*.

platform (Illumina Inc., CA, USA) with 150 bp paired-end reads, producing 81 M reads (Table 1). After chromatin crosslinking, the sample underwent nuclease digestion, fragmenting the genomic DNA while maintaining spatial proximity between interacting regions. Chromatin fragments were captured using magnetic beads targeting DNA–protein complexes. In the proximity ligation step, adjacent DNA fragments were ligated to create junctions that reflect their three-dimensional organization within the nucleus. The resulting DNA was purified and processed into sequencing libraries, including end repair and adapter ligation. The insert size distribution was evaluated using a Bioanalyzer (Agilent Technologies, CA, USA).

Genome assembly. Nanopore long reads were assembled with Hifiasm v0.25.0-r726[6], redundant contigs were removed with purge_dups v1.2.6[7] and scaffolded with Omni-C data using SALSA v2.3[8] with the "-e DNASE" option. Hifiasm was run in ONT mode using the option "-ont". The final assembly comprised 929 scaffolds totaling 2.16 Gb. An Omni-C contact map was generated with Juicer v1.6[9] and Juicebox v3.1.4[10] to validate chromosome-scale scaffolding. Assembly completeness was assessed using BUSCO v5.8.3[11] with the mammalia_odb10 dataset (n = 9226), yielding C:98.2% [S:94.3%, D:4.0%], F:0.7%, M:1.0% (Table 2). Completeness was assessed using BUSCO v5.8.3 with the mammalia_odb10 dataset for both the genome assembly and the predicted CDS set.

Gene prediction and annotation. Gene prediction was performed with Helixer v0.3.5[12], which identified 20,647 coding sequences (CDSs). The predicted gene set had a BUSCO score of C:92.2% [S:88.9%, D:3.3%], F:3.3%, M:4.5% (Table 2). The predicted CDS sequences were then compared against the black flying fox reference protein dataset (*Pteropus alecto*; GCA_000325575) using BLASTX, and functional domains were annotated using InterProScan v5.76–107.0[13]. In total, 19029 CDSs (92.1%) had BLASTX hits (E-value ≤ 1e-6), and 17407 CDSs (84.3%) were assigned at least one InterPro entry. A total of 1333 CDSs (6.4%) were not annotated by either method. Summary statistics of predicted gene structures (transcript length, CDS length, exon/intron counts and lengths) are provided in Supplementary File.

Phylogenomics. Phylogenomic analysis based on BUSCO single-copy orthologs was performed to clarify the evolutionary placement of *P. pselaphon* within the *Pteropus* bats. Amino acid sequences of orthologs from *P. rufus* (GCA_028533765), *P. medius* (GCA_902729225), *P. vampyrus* (GCA_000151845), and *P. alecto* (GCA_000325575) were included, with *Rousettus aegyptiacus* (GCA_014176215), *Rhinolophus ferrumequinum* (GCA_004115265), and *Hipposideros armiger* (GCA_001890085) as outgroups. Using the BUSCO_phylogenomics pipeline[14] (v20240919), amino acid sequences of shared single-copy orthologs were extracted, aligned with MUSCLE, and trimmed with trimAl using default settings as implemented in BUSCO_phylogenomics. We retained only loci present in all taxa for supermatrix construction (locus occupancy = 100%), resulting in a total of 7,259 shared single-copy orthologs. The trimmed alignments were concatenated into a single amino-acid supermatrix without partitioning. The resulting supermatrix (8 taxa; 4,756,266 sites) showed high completeness, with an overall site occupancy of 95.0% (proportion of non-gap characters across taxa). A genome-scale maximum-likelihood phylogeny was inferred with IQ-TREE2[15] (v2.4) using a Q.mammal + F + I + R10 substitution model selected by ModelFinder module. Phylogenetic inference and nodal support assessment were performed in IQ-TREE2, and node labels in Fig. 3 indicate ultrafast bootstrap (UFBoot) support values based on 1,000 replicates. The resulting tree was visualized and annotated with FigTree v1.4.4[16]. The topology supported *P. pselaphon* as a distinct lineage within the *Pteropus* clade (Fig. 3).

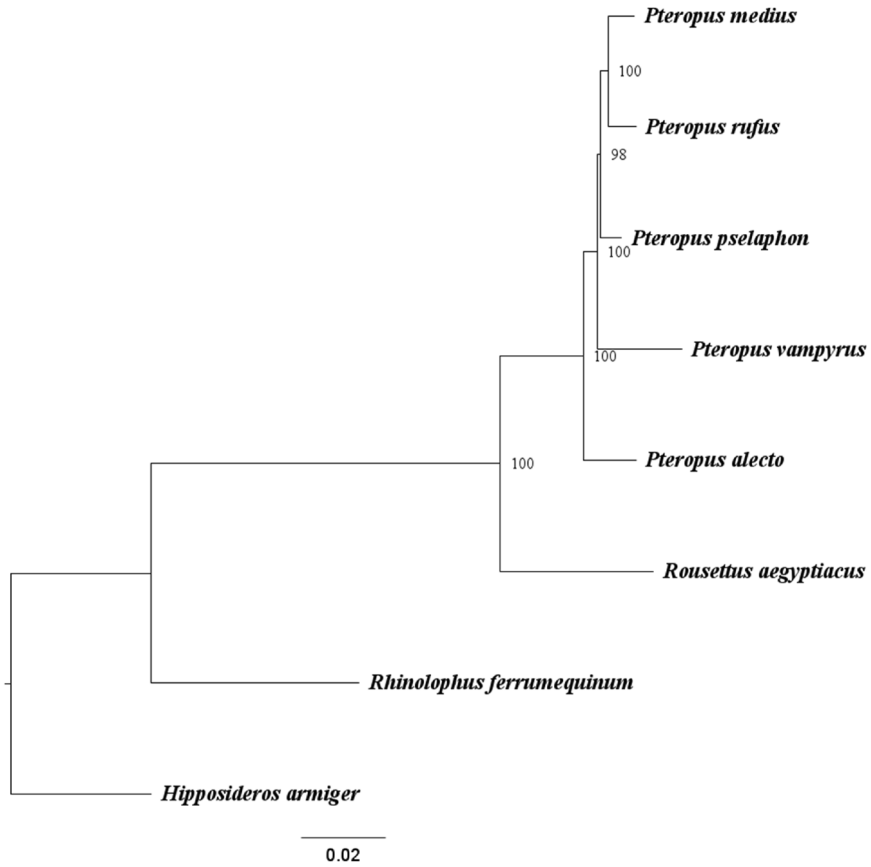

Fig. 3 A genome scale phylogenetic tree of genus *Pteropus*. Maximum-likelihood tree inferred from a concatenated amino-acid supermatrix of 7,259 BUSCO single-copy orthologs using IQ-TREE2 under the Q.mammal + F + I + R10 model. Numbers on nodes indicate ultrafast bootstrap support values (UFBoot, %, 1,000 replicates). Branch lengths are measured in substitutions per site; the scale bar indicates substitutions per site.

Transposable element annotation. We identified and masked transposable elements (TE) in the *Pteropus pselaphon* genome using RepeatModeler2 v2.0.7[17] and RepeatMasker v4.2.1[18]. First, we built a de novo repeat library from the soft-masked primary assembly using RepeatModeler2 with default parameters. The resulting species-specific consensus library was then used as input for RepeatMasker using only the de novo consensus library generated by RepeatModeler2. We summarized the genomic proportion of each major repeat class (LINE, SINE, LTR elements, DNA transposons, and unclassified repeats) from the RepeatMasker output to characterize the repeat landscape of the *P. pselaphon* genome.

Data Records

The genome assembly of *Pteropus pselaphon* is publicly available at GenBank/NCBI Assembly under accession GCA_056242155.1[19], linked to BioProject PRJDB39851. The genome annotation file (GFF) is publicly available as part of the GenBank assembly record. Raw sequencing data are available from the DDBJ Sequence Read Archive under the INSDC study accession DRP016843 (run accessions: DRR892950 and DRR888616)[20]. The genome assembly is deposited in the DDBJ/ENA/GenBank INSDC database as a WGS project (BAAJCM010000000.1; WGS set BAAJCM010000000)[21].

Data Overview

RepeatMasker analysis showed that repetitive elements occupy 632.4 Mb of the *Pteropus pselaphon* genome, corresponding to 29.17% of the assembly (Fig. 4). Interspersed repeats account for 27.20% (589.6 Mb), with retroelements representing the most abundant class (14.18%), dominated by long interspersed nuclear elements (LINEs; 11.95%) compared with short interspersed nuclear elements (SINEs; 1.26%) and long terminal repeat (LTR) elements (0.97%). In addition, unclassified interspersed repeats constitute 12.28% of the genome, suggesting the presence of lineage-specific or poorly characterized repeat families. Small RNA-related repeats, simple repeats, and low-complexity regions contribute 0.29%, 1.45%, and 0.24% of the genome, respectively.

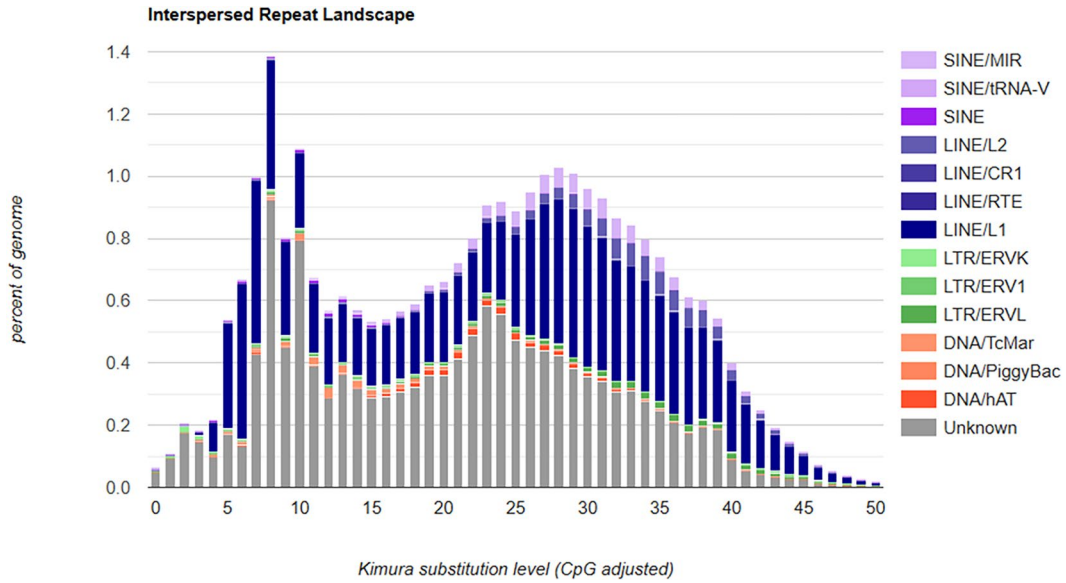

Fig. 4 Transposable element landscape of the *Pteropus pselaphon* genome. Transposable element (TE) landscape of the *P. pselaphon* genome. The x-axis shows the Kimura substitution level from TE consensus sequences, and the y-axis shows the genomic proportion occupied by each TE class or family.

	***P. pselaphon* (BAAJCM010000000)**	***P. rufus* (GCA_028533765)**	***P. medius* (GCA_902729225)**	***P. vampyrus* (GCA_000151845)**	***P. Alecto* (GCA_000325575)**
Number of scaffolds*	929	469091	10157	36094	21316
Total length	2,167,613,222	2,094,173,032	1,982,912,732	2,198,284,804	1,973,640,693
N50	81,982,708	110,476,781	18,871,253	5,954,017	16,439,519
L50	11	8	32	102	35
GC-content	40.29	39.77	39.82	39.81	39.71
BUSCO	C:98.2% [S:94.3%, D:4.0%], F:0.7%, M:1.0%	C:96.1% [S:95.9%, D:0.2%], F:1.7%, M:2.1%	C:97.3% [S:97.0%, D:0.3%], F:1.1%, M:1.6%	C:94.2% [S:93.8%, D:0.4%], F:2.9%, M:2.9%	C:97.5% [S:97.4%, D:0.1%], F:1.1%, M:1.3%

Table 3. Evaluation of the Genome Sequence of the genus *Pteropus*. *Including unplaced contigs.

Technical Validation

The *P. pselaphon* assembly demonstrated high continuity and completeness (BUSCO C:98.2%; scaffold N50: 82.0 Mb; Table 3). Assembly statistics (including scaffold N50/L50) were obtained from the corresponding GenBank assembly records, and BUSCO scores were computed in this study using BUSCO v5.8.3 (mammalia_odb10) on each assembly. While the scaffold N50 is lower than that of *P. rufus*, it exceeds those of *P. medius, P. vampyrus* and *P. alecto*, and the BUSCO completeness is the highest among the compared assemblies (Table 3). Omni-C contact maps generated with Juicer exhibited clear linear interaction patterns, confirming accurate chromosome-scale scaffolding (Fig. 5). Oxford Nanopore long reads were mapped back to the assembly using minimap2[22] (v2.30), resulting in 99.6% of reads aligning to the genome. Omni-C Illumina reads were mapped with BWA-MEM2[23] (v2.3), with 98.22% of read pairs mapping. The assembly contains 149 scaffold gaps. The 20 longest scaffolds account for ~77% of the total assembly length, supporting chromosome-scale scaffolding, while the remaining sequence (~23%) remains in 909 smaller unplaced scaffolds. Potential contamination by *Staphylococcus aureus* associated with the MRSA infection was assessed using Kraken2[24] (v2.1.3) on the genome. No scaffolds with substantial length (>10 kb) were classified as bacterial, indicating negligible contamination by MRSA-derived sequences.

Data availability

All data are publicly available from DDBJ/ENA/GenBank under BioProject PRJDB39851. The genome assembly generated in this study is available at GenBank under accession GCA_056242155.1. The genome annotation file (GFF) is publicly available as part of the GenBank assembly record. Raw sequencing data are available in the DDBJ Sequence Read Archive under accession DRR892950 (Nanopore long reads) and DRR888616 (Omni-C Illumina reads). The genome assembly is available as a Whole Genome Shotgun (WGS) project under accession BAAJCM010000000 (master record: BAAJCM000000000.1). Assembly: https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_056242155.1. BioProject: https://www.ncbi.nlm.nih.gov/bioproject/PRJDB39851/. WGS master record: https://www.ncbi.nlm.nih.gov/nuccore/BAAJCM000000000.1.

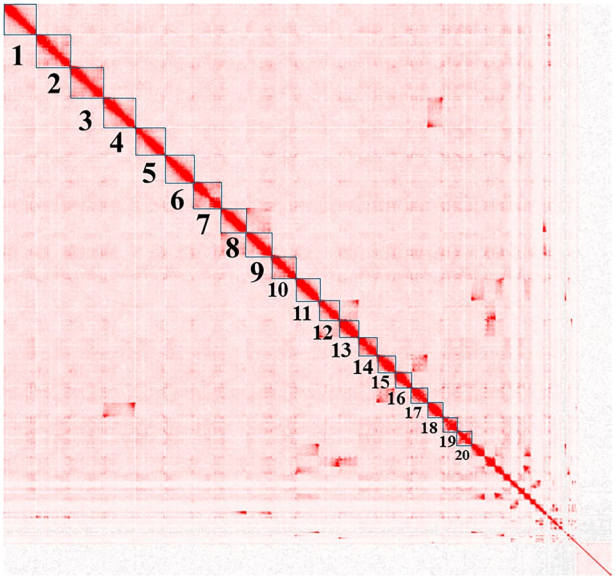

Fig. 5 Chromatin interaction heatmap of the *Pteropus pselaphon* genome. Omni-C contact map of the *Pteropus pselaphon* genome assembly showing chromatin contact frequencies among all scaffolds. Scaffold boundaries are indicated by blue squares, and the 20 largest scaffolds, ordered by scaffold length, are numbered along the axes for reference.

Code availability

No custom code was developed for this study. All software was used according to the manuals and protocols provided by the software developers. If no detailed parameters are mentioned for a given program, it was run using its default parameter settings, as described in the corresponding software documentation.

Received: 26 December 2025; Accepted: 13 April 2026;
Published online: 20 April 2026

References

1. Vincenot, C. E. Pteropus pselaphon. The IUCN Red List of Threatened Species 2017: e.T18752A22085351. Available at https://doi.org/10.2305/IUCN.UK.2017-2.RLTS.T18752A22085351.en. (accessed 28 November 2025) (2017).
2. Japan. Act on Conservation of Endangered Species of Wild Fauna and Flora (Act No. 75 of 1992). Japanese Law Translation https://www.japaneselawtranslation.go.jp/en/laws/view/2103/en (last version shown on page: Act No. 37 of 2019).
3. Ministry of the environment, Japan. Conservation Status of the Bonin Flying Fox. *(in Japanese)*. https://www.env.go.jp/nature/kisho/hogozoushoku/ogasawaraokomori.html.
4. Okada, A., Suzuki, H., Inaba, M., Horikoshi, K. & Shindo, J. Genetic structure and cryptic genealogy of the Bonin flying fox Pteropus pselaphon revealed by mitochondrial DNA and microsatellite markers. *Acta Chiropterologica* **16**, 15–26, https://doi.org/10.3161/150811014X683237 (2014).
5. De Coster, W. & Rademakers, R. NanoPack2: population-scale evaluation of long-read sequencing data. *Bioinformatics* **39**, btad311, https://doi.org/10.1093/bioinformatics/btad311 (2023).
6. Cheng, H., Asri, M., Lucas, J., Koren, S. & Li, H. Scalable telomere-to-telomere assembly for diploid and polyploid genomes with double graph. *Nat. Methods* **21**, 967–970, https://doi.org/10.1038/s41592-024-02269-8 (2024).
7. Guan, D. *et al*. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898, https://doi.org/10.1093/bioinformatics/btaa025 (2020).
8. Ghurye, J. *et al*. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS Comput. Biol.* **15**, e1007273, https://doi.org/10.1371/journal.pcbi.1007273 (2019).
9. Durand, N. C. *et al*. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98, https://doi.org/10.1016/j.cels.2016.07.002 (2016).
10. Durand, N. C. *et al*. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101, https://doi.org/10.1016/j.cels.2015.07.012 (2016).
11. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212, https://doi.org/10.1093/bioinformatics/btv351 (2015).
12. Holst, F. *et al*. Helixer: ab initio prediction of primary eukaryotic gene models combining deep learning and a hidden Markov model. *Nat. Methods* https://doi.org/10.1038/s41592-025-02939-1 (2025).
13. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240, https://doi.org/10.1093/bioinformatics/btu031 (2014).
14. McGowan, J. BUSCO_phylogenomics (v20240919). *GitHub repository* https://github.com/jamiemcg/BUSCO_phylogenomics (2024).
15. Minh, B. Q. *et al*. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534, https://doi.org/10.1093/molbev/msaa015 (2020).
16. Rambaut, A. FigTree v1.4.4. Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK. http://tree.bio.ed.ac.uk/software/figtree/ (2018).
17. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl Acad. Sci. USA* **117**, 9451–9457, https://doi.org/10.1073/pnas.1921046117 (2020).

18. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4.0. http://www.repeatmasker.org (2013–2015).
19. Nabeshima, K., Shimada, Y. & Onuma, M. *GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_056242155.1 (2026).
20. *NCBI Sequence Read Archive*. https://identifiers.org/ncbi/insdc.sra:DRP016843 (2025).
21. *INSDC WGS project*. https://identifiers.org/ncbi/insdc:BAAJCM000000000.1 (2025).
22. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100, https://doi.org/10.1093/bioinformatics/bty191 (2018).
23. Vasimuddin, M., Misra, S., Li, H. & Aluru, S. Efficient architecture-aware acceleration of BWA-MEM for multicore systems. In *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)* 314–324, https://doi.org/10.1109/IPDPS.2019.00041 (IEEE, 2019).
24. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46, https://doi.org/10.1186/gb-2014-15-3-r46 (2014).

Acknowledgements

We sincerely thank the Ministry of the Environment Ogasawara Ranger Office for Conservation, the Institute of Boninology, Council for Island Development through Coexistence of People, Pets and Wild Animals in Ogasawara, and the Ueno Zoological Gardens, Tokyo Zoological Park Society, for providing valuable support and sample access for this study. This research received no external funding and was supported by institutional resources of the National Institute for Environmental Studies and Mie University.

Author contributions

K.N. conceived the study, performed the main analyses, and wrote the main manuscript text. Y.S. performed the sequencing and revised the manuscript. M.O. supervised the study and provided funding support. All authors reviewed and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-026-07245-9.

Correspondence and requests for materials should be addressed to K.N.