scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A chromosome-level genome assembly of the Leschenault's rousette (*Rousettus leschenaultii*)

Lijia Chen[1,2], Guang Yang[2,3], Shengwei Hou[1] ✉ & Simin Chai[2] ✉

The Leschenault's rousette (*Rousettus leschenaultii*) represents a medium-sized bat of the genus *Rousettus* that feeds on fruits or flowers, mainly distributed in Southeast Asia. Here, we generated a chromosome-level genome assembly of *R. leschenaultii* using a hybrid MGI and PacBio sequencing approach, facilitated by the chromosome assignment using the high-throughput chromatin conformation capture (Hi-C) sequencing technology. The genome size was 1.95 Gb with a scaffold N50 of 116.99 Mb, and 99.00% of the assembled sequences were anchored to 17 autosomes and the two sex chromosomes (X and Y). The completeness of the assembly was estimated to be 96.4% using BUSCO. In total, 19,625 genes were predicted from this genome assembly, with 97.83% of them being functionally annotated. This high-quality assembly of *R. leschenaultii* serves as a valuable genetic resource for exploring the genomic basis of evolutionary adaptation, and for conducting population, ecological, and conservation genomic studies.

Background & Summary

The Leschenault's rousette (*Rousettus leschenaultii*) is a tropical cave-dwelling species that can survive in a temperate climate, and it inhabits a range of environments, from tropical forests to urban areas[1]. A colony of these bats can contain up to several thousand individuals. In a previous study, the population size of this species was assessed as "least concern" (LC), showing a stable population trend as reported by the International Union for Conservation of Nature (IUCN)[2]. However, a subsequent assessment by the IUCN[3] has identified a decline in the population size of *R. leschenaultii*, prompting a reassessment of the species as near threatened (NT). It is of paramount importance to maintain a stable population size and to investigate the genetic bias underlying the intriguing traits of *R. leschenaultii*.

R. leschenaultii is an informative diploid mammalian model for studying adaptive evolution. It exhibits pronounced ecological versatility—forming large cave colonies, foraging across forest–urban mosaics, and contributing to seed dispersal[4,5]. It further displays rapid postnatal wing growth and tongue-click echolocation, which is uncommon in bats and unique within the family Pteropodidae[6,7]. Early population genetic analyses based on mitochondrial DNA suggested high gene flow and little substructure across populations[8]. In addition, comparative genomic analyses of olfactory and trace-amine receptor families and immune pathways reveal signatures of lineage-specific or convergent adaptation[9,10]. To conduct rigorous analyses of structural and copy-number variation and genome-wide selection scans, high-quality, chromosome-level reference genomes are indispensable. Although a chromosome-level genome assembly has recently been available for *R. leschenaultii*[11], additional high-quality reference genome resources remain valuable for rigorous evolutionary and population-genomic analyses in broader comparative contexts.

In this study, we employed a hybrid sequencing approach combining MGI short-read sequencing (65.20 Gb) and PacBio HiFi long-read sequencing (91.54 Gb) for assembly, and used the Hi-C technology (61.63 Gb) for chromosomal assignment (Table 1). *R. leschenaultii* is a diploid mammalian species, and the genome presented here represents a haploid-sized primary assembly derived from a diploid individual, resulting in a chromosome-level assembly of *R. leschenaultii* with a total length of 1.95 Gb and a scaffold N50 of 116.99 Mb.

[1]Department of Ocean Science and Engineering, Southern University of Science and Technology, Shenzhen, 518055, China. [2]Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou), Guangzhou, 511458, China. [3]Jiangsu Key Laboratory for the Biodiversity Conservation and Sustainable Utilization in the Middle and Lower Reaches of the Yangtze River Basin, College of Life Sciences, Nanjing Normal University, Nanjing, 210023, China. ✉e-mail: housw@sustech.edu.cn; chaisiminendeavor@163.com

Library	Read length/ size selection (bp)	Reads number	Raw data (Gb)	Coverage (×)
MGI	150	434,651,852	65.20	33.4
PacBio	15,000	4,967,715	91.54	46.9
Hi-C	150	410,859,976	61.63	31.6

Table 1. Sequencing data and methods used in this study to assemble the *R. leschenaultii* genome.

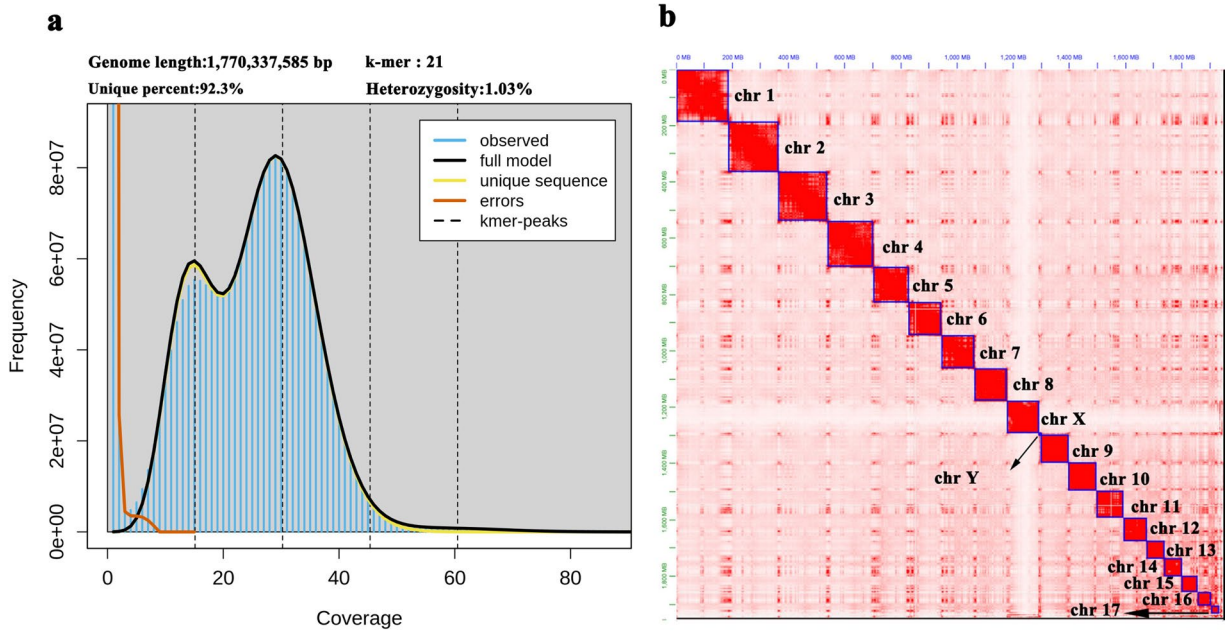

Fig. 1 Overview of the genome of *R. leschenaultii*. (**a**) The k-mer distribution of MGI paired-end reads inferred by GenomeScope based on the 21-mer. (**b**) Genome-wide Hi-C heatmap of *R. leschenaultii*; the blue squares represent chromosomes.

Our assembly offered an important genomic asset for evolutionary and population genomic studies of *R. leschenaultii*.

Methods

Sample collection and sequencing. A naturally deceased specimen of *R. leschenaultii* was obtained from a cave in Dali City, Yunnan Province. The deep thigh muscle of the individual was promptly frozen in liquid nitrogen for preservation, serving as the sequencing tissue. DNA samples for short-read sequencing and long-read sequencing were extracted from the muscle tissue of *R. leschenaultii* using the CTAB method, and DNA libraries were constructed by MGIEasy Universal DNA Library Prep Kit V1.0 (CAT#1000005250, MGI) following the standard protocol. Then, qualified libraries were sequenced on the DNBSEQ-T7 platform with paired-end 150 bp reads in GrandOmics Bioscience Co., Ltd. As a result, a total of 65.20 Gb of short reads were obtained.

For PacBio long-read sequencing, SMRTbell libraries were constructed according to PacBio's standard protocol (Pacific Biosciences, CA, USA) using 15 kb size-selected DNA preparation. Sequencing was performed on a PacBio Revio instrument with Sequencing Primer V5 and Revio Binding Kit 2.2 in GrandOmics Bioscience Co, Ltd, generating a total of 91.5 Gb raw data, with the maximum read length of 57.90 kb and the N50 of 19.15 kb.

For the Hi-C sequencing, muscle tissue was used for genomic DNA sequencing. Cells were cross-linked with 1% formaldehyde, and chromatin was digested with the restriction enzyme DpnII. Proximity ligation was then performed to ligate spatially proximal DNA fragments within cross-linked chromatin[12]. After reversal of cross-links, biotin-labelled DNA fragments were enriched and used for library construction with the NEBNext Ultra II DNA Library Prep Kit for Illumina (NEB) according to the manufacturer's instructions. The final library was sequenced on the DNBSEQ-T7 platform with paired-end 150 bp reads, resulting in the generation of 61.63 Gb raw data.

Genome survey. To obtain the main characteristics of the genome, including estimated genome size, duplicated sequence content, and heterozygosity through a genome survey, MGI short reads (65.2 Gb) were assessed and trimmed by Fastp v0.23.4[13] with default parameters. Then, the genome survey was conducted by Jellyfish v2.2.10[14] and GenomeScope v2.0[15] based on the 21-mer distribution. The estimated genome size of *R. leschenaultii* was 1.77 Gb with a duplicated content of 39.20% and a heterozygosity of 1.03% (Fig. 1a).

Feature	Value
Estimated genome size (bp)	1,770,337,585
Total length (bp)	1,952,227,642
Contigs number	109
Contigs N50 (bp)	101,034,847
Scaffold number	74
Scaffold N50 (bp)	116,991,804
GC content (%)	40.37
Total base on chromosomes (bp)	1,932,802,446
Number of gaps (consecutive run of "N"s)	35

Table 2. Estimates of the basic genomic features of *R. leschenaultii*.

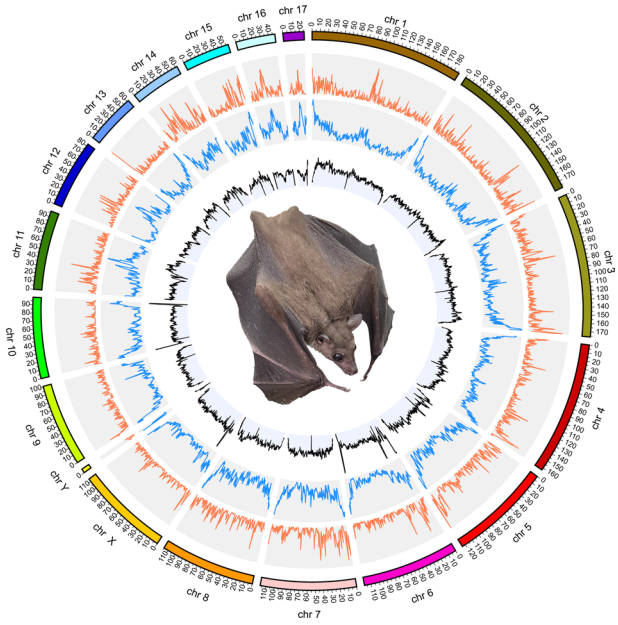

Fig. 2 Circos plot illustrates the genomic structural characteristics of *R. leschenaultii* assembly in the present study. Tracks from outer to inner display information for chromosomes, gene density, GC content, and repeat density, respectively.

Genome assembly. The contig-level assembly was obtained using hifiasm v0.19.8-r603[16] with default parameters based on the PacBio HiFi long reads. Redundant fragments in the contig assembly were removed by Purge_dups v1.2.6[17]. To obtain the chromosome-level assembly, Hi-C data were used to further anchor the assembly. Specifically, the contig sequences were indexed by SAMtools v1.16[18], Hi-C clean reads were mapped to the contig assembly using BWA[19], and duplicated reads were marked by picard[20]. A nearly chromosome-level assembly was conducted by Juicer v1.2[21] and YaHS v1.2[22]. Subsequently, Juicebox v2.17.00[23] was employed to manually correct the chromatin contact matrix, and the final assembly was recombined by Juicer v1.2. The heatmap was drawn to illustrate the interaction of each chromosome (Fig. 1b). Consequently, the final assembly has a genome size of 1.95 Gb and scaffold N50 of 116.99 Mb (Table 2). The genomic characteristics of the assembled genome are visualized in Circos plots, showing gene density, GC content, and repeat density (Fig. 2).

Whole-genome alignment. Whole-genome alignment analysis was performed based on the *R. leschenaultii* genome assembled in this study, a previously published *R. leschenaultii* assembly[11], and an assembly from the closely related species *R. aegyptiacus*[24]. Sequence homology was identified with MUMmer4[25] and then visualized using NGenomeSyn[26] (Fig. 3). The chromosome X was identified based on extensive conserved homology with the chromosome X of *R. aegyptiacus*. Given the high interspecific variation of the Y chromosome[27], the blastn[28] was therefore used to validate the assembly of the Y chromosome by aligning with Y-linked genes in mammals (*USP9Y* and *UTY*, Supplementary Table 1)[29]. Together, the highly matched homology and the Y-linked gene placements support the accuracy of our assembly.

Repeats prediction. Both *de novo* and homology-based prediction methods were employed to annotate transposable elements (TEs) of the *R. leschenaultii* genome. Briefly, a *de novo* repeat library was constructed

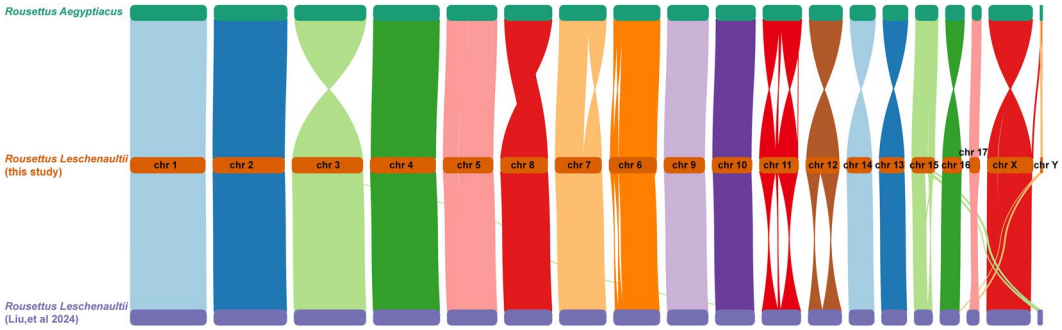

Fig. 3 Whole-genome alignment and sequence homology among assemblies of *R. aegyptiacus* and *R. leschenaultii*. Chromosomes for *R. leschenaultii* genome assembly from the present study are explicitly labeled.

Type	Dfam TEs		Protein TEs		*De novo* TEs		Combined TEs	
	Length (bp)	% of genome	Length (bp)	% of genome	Length (bp)	% of genome	Length (bp)	% of genome
DNA	71,036,879	3.638	2,038,976	0.104	47,512,604	2.434	74,673,482	3.825
LINEs	352,070,466	18.033	125,564,073	6.431	288,626,555	14.783	395,954,545	20.28
SINEs	66,434,017	3.403	0	0	39,652,863	2.031	73,452,089	3.762
LTRs	121,478,682	6.222	6,937,738	0.355	83,807,509	4.293	128,212,337	6.567
Unknown	1,408,296	0.072	0	0	5,332,145	0.273	6,655,070	0.341
Total	665,823,786	34.103	134,535,352	6.891	468,686,903	24.006	690,199,901	35.351

Table 3. Statistics of repeat elements of the *R. leschenaultii* genome. Dfam TEs and Protein TEs were identified using homology-based methods, whereas *de novo* TEs were predicted using *de novo* prediction methods.

Database	Number	Percent (%)
eggNOG	18,076	92.11
Pfam	18,015	91.80
Swiss-Prot	19,029	96.96
NR	19,195	97.81
Merge	19,199	97.83

Table 4. Statistics of gene functional annotation in the *R. leschenaultii* genome.

using RepeatModeler v2.0.3[30]. The homology method was performed using RepeatMasker v4.1.3[31] against the Dfam database and species-specific repeat library. Dfam TEs and *De novo* TEs denote RepeatMasker annotations based on the Dfam library and the RepeatModeler-derived *de novo* library, respectively. Protein TEs were identified by annotating the genome with the RepBase protein data[32] using RepeatProteinMask[31]. Correspondingly, in Table 3, Dfam TEs and Protein TEs were obtained using homology-based approaches, whereas *de novo* TEs were predicted through *de novo* repeat discovery; Combined TEs represent the results obtained by integrating the repetitive element annotations above and removing redundancy. A total of 38.23% of the genome was identified as repetitive elements, of which transposable elements accounted for 35.35% of the genome. Specifically, predicted TEs contain DNA transposons (3.83%), LINEs (20.28%), SINEs (3.76%), LTRs (6.57%), and unknown (0.34%) (Table 3).

Gene prediction and functional annotation. *De novo* prediction and homology-based methods were combined to perform gene prediction after masking the repeat sequences in *R. leschenaultii*. The *de novo* approach was conducted by Augustus v3.4.0[33], GlimmerHMM[34], and GENSCAN[35]. For the homology-based gene prediction, the published genome sequence and annotation of seven species, including four bats (*Desmodus rotundus* (GCF_022682495.1), *Pteropus vampyrus* (GCF_902729225.1), *Rhinolophus ferrumequinum* (GCF_004115265.2), *Phyllostomus discolor* (GCF_004126475.2)) and three other mammals (*Homo sapiens* (GCF_000001405.40), *Canis lupus familiaris* (GCF_011100685.1), *Equus caballus* (GCF_002863925.1)) from NCBI GenBank were used for aligning the protein sequences of *R. leschenaultii* using GeMoMa[36]. The results of these two approaches were merged into a consensus gene set by using EvidenceModeler v2[37]. Eventually, 19,625 genes were predicted in the genome. Furthermore, the functional annotation was conducted by alignment to the NR[38], Swiss-Prot[39], Pfam[40], and eggNOG[41] databases. A total of 19,199 genes (97.83%) were annotated (Table 4), and the results of functional annotation were visualized by a Venn[42] plot (Fig. 4).

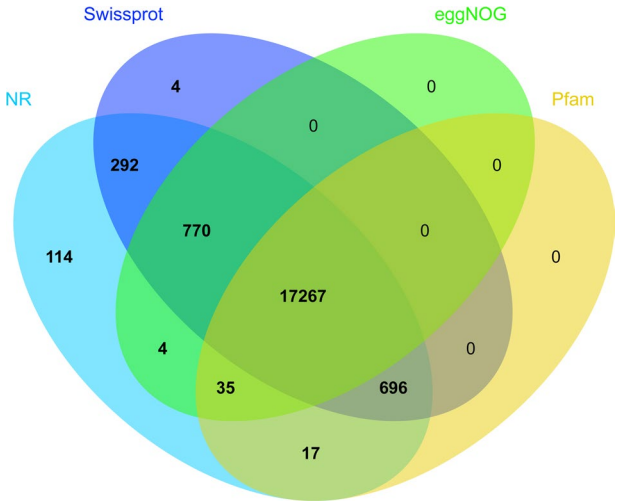

Fig. 4 Venn diagram of protein-coding genes annotated against different databases.

Type	Genome assembly		Predicted genes	
	Number	Percentage (%)	Number	Percentage (%)
Complete BUSCOs	8894	96.4	9061	98.2
Single-Copy BUSCOs	8873	96.2	9045	98.0
Duplicated BUSCOs	21	0.2	16	0.2
Fragmented BUSCOs	68	0.7	5	0.1
Missing BUSCOs	264	2.9	160	1.7
Short-read mapping rate	—	99.92	—	—
PacBio long-read mapping rate	—	99.99	—	—

Table 5. Summary of read mapping rates and BUSCO results for the *R. leschenaultii* genome assembly and predicted genes.

Data Records

The sequencing reads generated in this study are available in the NCBI Sequence Read Archive (SRA) under BioProject accession PRJNA1335844[43], with run accessions SRR35761212[44] (MGI sequencing), SRR35761210[45] (Hi-C sequencing), and SRR35761211[46] (PacBio sequencing). The genome sequence has been deposited in GenBank under the accession number JBRKBD000000000[47], and the corresponding annotation files are available on Figshare[48].

Technical Validation

The completeness and accuracy of the assembly of *R. leschenaultii* were evaluated by multiple methods. The Benchmarking Universal Single-Copy Orthologs (BUSCO v5.4.7)[49] was used to evaluate the completeness of both the genome and predicted gene set against the mammalia_odb10 database, yielding completeness scores of 96.4% and 98.2%, respectively. Base-pair accuracy was further assessed using Merqury v1.3[50], which produced a consensus quality value (QV) of 47.95. In addition, the mapping rates of MGI short reads and PacBio long reads were 99.92% and 99.99%, respectively, suggesting a high recruitment of generated reads (Table 5). Compared with the chromosome-level assembly reported by Liu *et al.*[11], our assembly demonstrates higher base-level accuracy and gene completeness (Table 6). Taken together, these results suggested that the assembled *R. leschenaultii* genome was of high quality at the chromosome level.

Ethics statement. No live animals were used or euthanized in this study. The specimen analyzed in this study was a naturally deceased individual of *Rousettus leschenaultii* encountered during fieldwork in Yunnan Province, China.

Data availability

Raw sequencing reads are available in the NCBI SRA under BioProject PRJNA1335844. This Whole Genome Shotgun project has been deposited at GenBank under the accession JBRKBD000000000 (version JBRKBD020000000), and the corresponding genome annotation files are available on Figshare (https://doi.org/10.6084/m9.figshare.30360511.v3).

Genome assembly	This study	Previously published assembly
Scaffold Number	74	161
Scaffold N50 (bp)	116,991,804	113,931,000
Maximum scaffold length (bp)	186,544,798	186,013,523
Total length (bp)	1,952,227,642	1,902,122,851
Complete BUSCO percentage (%)	96.4	93.4
Fragmented BUSCO percentage (%)	0.7	1.8
Missing BUSCO percentage (%)	2.9	4.8
QV	47.95	40.45

Table 6. Comparative summary of our newly assembled *R. leschenaultii* genome in the present study and a previously published *R. leschenaultii* assembly.

Code availability

No special code or scripts were used in this work, and data processing was carried out based on the protocols and manuals of the corresponding bioinformatics software.

Received: 25 October 2025; Accepted: 9 March 2026;
Published online: 13 March 2026

References

1. Acharya, P. & Adhikari, H. *Bats of Nepal: A Field Guide* (2010).
2. Bates, P.H. & Rousettus leschenaultii, K. The IUCN Red List of Threatened Species 2008 (2008).
3. Bouillard, N., Cretan, S. & Waldien, D. *Rousettus leschenaultii: IUCN Red List of Threatened Species 2021* (2021).
4. Raghuram, H., Thangadurai, C., Gopukumar, N., Nathar, K. & Sripathi, K. The role of olfaction and vision in the foraging behaviour of an echolocating megachiropteran fruit bat, Rousettus leschenaulti (Pteropodidae). *Mammalian Biology* **74**, 9–14 (2009).
5. Tang, Z.-H. *et al*. Seed dispersal of Syzygium oblatum (Myrtaceae) by two species of fruit bat (Cynopterus sphinx and Rousettus leschenaulti) in South-West China. *Journal of Tropical Ecology* **28**, 255–261 (2012).
6. Elangovan, V., Raghuram, H., Satya Priya, E. Y. & Marimuthu, G. Wing Morphology and Flight Performance in Rousettus leschenaulti. *Journal of Mammalogy* **85**, 806–812 (2004).
7. Yovel, Y., Geva-Sagiv, M. & Ulanovsky, N. Click-based echolocation in bats: not so primitive after all. *Journal of Comparative Physiology A* **197**, 515–530 (2011).
8. Chen, J. *et al*. Contrasting genetic structure in two co-distributed species of old world fruit bat. *PLoS One* **5**, e13903 (2010).
9. Nikaido, M. *et al*. Comparative genomic analyses illuminate the distinct evolution of megabats within Chiroptera. *DNA Res* **27** (2020).
10. Jebb, D. *et al*. Six reference-quality genomes reveal evolution of bat adaptations. *Nature* **583**, 578–584 (2020).
11. Liu, M. *et al*. Complexin-1 enhances ultrasound neurotransmission in the mammalian auditory pathway. *Nat Genet* **56**, 1503–1515 (2024).
12. Gong, G. *et al*. Chromosomal-level assembly of yellow catfish genome using third-generation DNA sequencing and Hi-C analysis. *GigaScience* **7** (2018).
13. Chen, S. F., Zhou, Y. Q., Chen, Y. R. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, 884–890 (2018).
14. Marcais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
15. Ranallo-Benavidez, T.R., Jaron, K.S. & Schatz, M.C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications* **11** (2020).
16. Cheng, H. Y., Concepcion, G. T., Feng, X. W., Zhang, H. W. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods* **18**, 170–+ (2021).
17. Guan, D. *et al*. Identifying and removing haplotypic duplication in primary genome assemblies. **36**, 2896–2898 (2019).
18. Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
19. Li, H.J.A.G. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM (2013).
20. Chong, R. A. & Mueller, R. L. Evolution along the mutation gradient in the dynamic mitochondrial genome of salamanders. *Genome Biol Evol* **5**, 1652–60 (2013).
21. Durand, N. C. *et al*. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems* **3**, 95–98 (2016).
22. Zhou, C.X., McCarthy, S.A. & Durbin, R. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics* **39** (2023).
23. Durand, N. C. *et al*. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Systems* **3**, 99–101 (2016).
24. Morales, A.E. *et al*. Bat genomes illuminate adaptations to viral tolerance and disease resistance. *Nature* **638** (2025).
25. Marçais, G. *et al*. MUMmer4: A fast and versatile genome alignment system. *Plos Computational Biology* **14** (2018).
26. He, W.M. *et al*. NGenomeSyn: an easy-to-use and flexible tool for publication-ready visualization of syntenic relationships across multiple genomes. *Bioinformatics* **39** (2023).
27. Li, G. *et al*. Comparative analysis of mammalian Y chromosomes illuminates ancestral structure and lineage-specific evolution. **23**, 1486-1495 (2013).
28. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic Local Alignment Search Tool. *Journal of Molecular Biology* **215**, 403–410 (1990).
29. Godfrey, A. K. *et al*. Quantitative analysis of Y-Chromosome gene expression across 36 human tissues. *Genome Res* **30**, 860–873 (2020).
30. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences of the United States of America* **117**, 9451–9457 (2020).
31. Tarailo-Graovac, S.M. & Chen, N. Using RepeatMasker to Identify Repetitive Elements in Genomic (2009).
32. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* **6**, 11 (2015).
33. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644 (2008).

34. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM:: two open source eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
35. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA11Edited by F. E. Cohen. *Journal of Molecular Biology* **268**, 78–94 (1997).
36. Keilwagen, J. *et al*. Using intron position conservation for homology-based gene prediction. *Nucleic Acids Research* **44** (2016).
37. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology* **9**, R7 (2008).
38. Marchler-Bauer, A. *et al*. CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Research* **39**, D225–D229 (2011).
39. Bairoch, A. *et al*. The universal protein resource (UniProt). *Nucleic Acids Research* **33**, D154–D159 (2005).
40. Mistry, J. *et al*. Pfam: The protein families database in 2021. *Nucleic Acids Research* **49**, D412–D419 (2021).
41. Huerta-Cepas, J. *et al*. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research* **47**, D309–D314 (2019).
42. Hittel, D. & Storey, K. B. The translation state of differentially expressed mRNAs in the hibernating 13-lined ground squirrel (Spermophilus tridecemlineatus). *Arch Biochem Biophys* **401**, 244–54 (2002).
43. *NCBI BioProject* https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1335844 (2025).
44. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35761212 (2025).
45. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35761210 (2025).
46. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35761211 (2025).
47. NCBI. *GenBank* https://identifiers.org/ncbi/insdc:JBRKBD000000000 (2025).
48. Chen, L. Genome annotation of Rousettus leschenaultii. *Figshare* https://doi.org/10.6084/m9.figshare.30360511.v3 (2025).
49. Manni, M., Berkeley, M. R., Seppey, M., Simao, F. A. & Zdobnov, E. M. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution* **38**, 4647–4654 (2021).
50. Rhie, A., Walenz, B.P., Koren, S. & Phillippy, A.M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology* **21** (2020).

Acknowledgements

We thank Dr. Xin Huang (Jiangxi Agricultural University) for technical guidance in bioinformatics for this study. We thank Prof. Tinglei Jiang (Northeast Normal University), Dr. Yujuan Wang (Northeast Normal University), Dr. Yuan Mu (Dali University), Ms. Yini Ren (Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou)) and Dr. Xiaobin Huang (Dali University) for sample collection and preparation. This study was supported by the Youth Fund of the National Natural Science Foundation of China (grant no. 32200345 to S.C.), the National Natural Science Foundation of China (grant no. 42476109 and 42276163 to S.H.), the Shenzhen Science, Technology and Innovation Commission Program (grant no. JCYJ20220530115401003 to S.H.), SUSTech Education Reform Programme (grant no. SJZLGC202437 to S.H.), and PI Project of Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou) (grant no. GML2021GD0805 to G.Y.).

Author contributions

G.Y., S.C., and S.H. designed and supervised the study. L.C. performed bioinformatics analysis and wrote the manuscript. S.C. collected the samples and wrote the manuscript. G.Y. and S.C. revised 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/s41597-026-07043-3.

Correspondence and requests for materials should be addressed to S.H. or S.C.