



OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of the Chinese soft-shelled turtle *Pelodiscus sinensis*

Jing Chen¹ , Jiayun Yao, Xuemei Yuan, Lei Huang, Xianqi Peng, Xialian Bu, Jinbiao Jiao & Haiqi Zhang[✉]

The Chinese soft-shelled turtle *Pelodiscus sinensis* is an economically important aquaculture species in Asia for its high nutritional and medicinal values. In recent years, with the continuous development of the *P. sinensis* breeding industry, the problems of germplasm resource degradation and population mixing have become increasingly prominent. In this study, a total of 471.77 Gb of sequencing data was generated, including 87.28 Gb (39.67 ×) of PacBio HiFi reads, 104.21 (47.36 ×) Gb of Illumina reads, 241.66 Gb (109.84 ×) of Hi-C data, and 38.62 Gb of RNA sequencing data. The final genome contained a length of 2.24 Gb with a contig N50 of 107.61 Mb and a scaffold N50 of 129.58 Mb. The final 2.14 Gb (95.42%) assembled genome sequences were anchored on 33 chromosomes, with a chromosome length that ranged from 13.04 Mb to 336.74 Mb. A total of 21,532 protein-coding genes were predicted and 21,149 genes were functionally annotated. The high-quality genome assembled in this study will represent a significant contribution to germplasm resources conservation of *P. sinensis*.

Background & Summary

The Chinese soft-shelled turtle (*Pelodiscus sinensis*) belongs to the order Testudines, family Trionychidae, and genus *Pelodiscus*, and is distributed in many Asian countries, including China, Japan, Korea, Vietnam, etc.¹. Due to its rich nutritional and medical values, the breeding industry of *P. sinensis* has developed rapidly in recent years. According to FAO data, the total production of *P. sinensis* in 2022 has reached 375,000 tons, making it one of the most important aquatic species². In China, previous studies classified *P. sinensis* populations into different strains based on their geographical distribution, including the northern strain from the northern region of Hebei province, the Yellow river strain from the Yellow river basin, the Dongting lake strain, Poyang lake strain, and Taihu lake strain from the Yangtze river basin, the southwestern strain from Guangxi province, Taiwan strains from southern and central Taiwan, etc.^{3,4}. With the expansion of aquaculture production, cross regional reproduction between different farms has led to the degradation of *P. sinensis* germplasm resources⁵. Furthermore, due to overfishing and non-standard introduction, the wild resources of *P. sinensis* have decreased⁶. It has been listed as a “vulnerable species” on the International Union for Conservation of Nature (IUCN) Red List of Endangered Species⁷.

At present, research on the evaluation of *P. sinensis* germplasm resources mainly focuses on morphological detection, mitochondrial diversity, and phylogenetic relationships between different strains^{4,8,9}. Moreover, the degree of genetic differentiation among different geographical populations of *P. sinensis* is still unclear. It was suggested that different habitats and a long evolutionary history might be the reasons for the genetic differentiation of *P. sinensis*³. With the development of sequencing technology, whole genome sequencing has largely overcome the limitations of traditional genetic methods such as the lack of molecular markers, providing a reference for germplasm resource conservation and genetic differentiation research^{10–12}. Although a genome of soft-shelled turtle has been published in 2013, this genome was a fragmented draft with scaffold N50 lengths of 3.33 Mb¹³. The high-quality reference genome of *P. sinensis* can promote and advance the conservation genetics and molecular mechanism research of important economic traits of this species.

This study applied a combination strategy of Illumina paired-end sequencing, PacBio HiFi, and High-throughput chromosome conformation capture (Hi-C) technologies to generate sequencing data for the

Agriculture Ministry Key Laboratory of Healthy Freshwater Aquaculture, Key Laboratory of Fish Health and Nutrition of Zhejiang Province, Key Laboratory of Fishery Environment and Aquatic Product Quality and Safety of Huzhou City, Zhejiang Institute of Freshwater Fisheries, Huzhou, 313001, China. ✉e-mail: zmk407@126.com

construction of the chromosome genome of *P. sinensis*. The total length of the genome is about 2.24 Gb, and more than 97.2% of the BUSCO genes were detected, with contig N50 lengths of 107.61 Mb, indicating excellent integrity and sequence continuity of the genome. A total of 21,532 protein coding genes were predicted in the assembled genome, with 98.22% of the genes successfully functionally annotated. In recent years, some genome research of turtle and tortoise species have been reported, including *Chelonia mydas*¹³, *Mauremys mutica*¹⁴, *Mauremys reevesii*¹⁵, *Rafetus swinhoei*¹⁶, *Gopherus agassizii*¹⁷, *Trachemys scripta elegans*¹⁸, *Platysternon megacephalum*¹⁹, *Chrysemys picta bellii*²⁰, *Aldabrachelys gigantea*²¹, *Pelochelys cantorii*²², etc. The high-quality chromosome level genomes provided in this study may further serve as a valuable resource for the evolutionary research of reptiles.

Methods

Sample collection and sequencing. A healthy 1-year-old female *P. sinensis* was collected from a breeding farm of Huzhou, Zhejiang Province, China (37.0750°N, 113.9221°E) in June 2022. Muscle, spleen, kidney, heart, lung, and liver tissues were collected from *P. sinensis*, and quickly frozen in liquid nitrogen for one hour and then stored at −80°C. Among them, liver tissue was used for DNA sequencing for genome assembly, while all tissues were used for RNA sequencing. Genomic DNA and RNA were extracted using the Genomic DNA Extraction Kit (Takara Bio Inc., Dalian, China) and RNAisoPlus Reagent (Takara Bio Inc., Dalian, China), respectively.

For short-read sequencing, the Illumina HiSeq X (Illumina, San Diego, CA, USA) was used to perform paired-end sequencing with an insert size of 350 bp. Moreover, fastp v 0.21.0 was used to evaluate the quality of raw reads with default parameters²³, and clean reads were obtained by removing reads containing adapter, low-quality and ploy-N. For long-read DNA sequencing, the PacBio HiFi sequencing was performed on a PacBio Sequel II platform with circular consensus sequencing (CCS) mode²⁴. To anchor scaffolds onto the chromosomes, a Hi-C library was constructed according to the protocol described previously^{25,26}. The liver tissue of *P. sinensis* was crosslinked using paraformaldehyde solution and enzymatically digested with MboI restriction enzyme. The ends of the restriction fragments were labeled with biotinylated nucleotides, and the ligated DNA was extracted, purified, and sheared into 350 bp fragments for Hi-C library construction. Finally, the library was quantified with Q-PCR method and sequenced with the Illumina HiSeq X platform (Illumina, San Diego, CA, USA). After removing adapters and low-quality short reads, a total of 241.66 Gb (109.84×) of Hi-C data was generated. In addition, total RNAs from the tissues of muscle, spleen, kidney, heart, lung, and liver tissues were extracted. Then, RNA quality and quantity of all tissues were detected by a NanoDrop spectrophotometer (NanoDrop products, Wilmington, DE, USA), a 2100 Bioanalyzer (Agilent Technologies, CA, USA), and 1% agarose gel electrophoresis. Finally, six RNA-seq library was constructed using the Illumina HiSeq X platform (Illumina, San Diego, CA, USA). Additionally, all tissues were equally mixed for Iso-Seq. The cDNA library was sequenced on the PacBio sequel II platform. In total, we obtained 471.77 Gb of sequencing data, which included 104.21 Gb (47.36×) of Illumina reads, 87.28 Gb (39.67×) of PacBio HiFi reads, 241.66 Gb (109.84×) of Hi-C data, and 38.62 Gb of RNA sequencing data.

De novo assembly and chromosome construction of the *P. sinensis* genome. The *k*-mer analysis was utilized to survey the genome features of *P. sinensis* with the Illumina short reads²⁷. Genome size, heterozygosity, and duplication rate were estimated using GenomeScope v 2.0²⁸. The 17-mer analysis estimated the genome size of *P. sinensis* was approximately 2.14 Mb, with a duplication rate of 52.49% and a heterozygosity of 0.81%. The initially assembly of PacBio HiFi long reads was generated using Hifiasm v 0.19.8 with the default parameters²⁹. The heterozygous sequences were removed using the Purge_haplotigs v 1.1.1 with default parameters³⁰. The draft genome contained a total size of 2.24 Gb containing 220 contigs with N50 sizes of 107.61 Mb. To assemble a chromosome-level genome, the Hi-C reads were mapped to the assembled genome and filtered by Jucier v 1.6³¹. The contigs were ordered and anchored into chromosomes using the 3D-DNA³², and manually adjusted using Juicebox³³. Finally, the Hi-C interaction heatmap demonstrated an excellent quality of the genome assembly (Fig. 1A). Approximately 805.56 million read pairs generated from Hi-C sequencing. Previous study revealed that *P. sinensis* had a diploid chromosome number of 33³⁴. The Circos³⁵ was used to visualize the 33 chromosomes, total TE density, DNA-TE density, LINE density, LTR density, and GC% density (Fig. 1B). The longest and shortest chromosomes were 336.74 Mb and 13.04 Mb in length, respectively (Table 1). For the final genome assembly, the contig N50 and scaffold N50 reached 107.61 Mb and 129.58 Mb, respectively (Table 2).

To evaluate the quality of the assembled genome, the completeness and accuracy of this genome were assessed by short-read mapping and Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis. Using BWA v0.7.10-r789³⁶, the short reads were aligned to the genome, it was found that over 98.43% of the reads were aligned, demonstrating a high mapping ratio for the short-read sequencing data. Furthermore, the completeness of the assembled *P. sinensis* genome was assessed by BUSCO v5.4.6 with the vertebrata_odb10 database³⁷. Among the 3354 single-copy orthologous genes, 3260 (97.2%) and 27 (0.8%) were identified as complete and fragmented BUSCOs, respectively, indicating that the assembled *P. sinensis* genome had high quality (Table 3).

Repetitive and non-coding gene prediction. The annotation of repetitive elements was divided into two methods: *de novo* prediction and homology-based alignment³⁸. In this study, repetitive elements and long terminal repeats were identified in the genome using RepeatModeler³⁹ and LTR-FINDER⁴⁰ with default parameters. Afterwards, the homology-based alignment was performed utilizing the RepBase database⁴¹. DNA and protein transposable elements (TEs) were detected by RepeatMasker and RepeatProteinMask⁴², respectively. Tandem repeats were identified with Tandem Repeat Finder⁴³. The repetitive element annotations are listed in Table 4. By combining Repbase and *de novo* datasets, we obtained a total of approximately 1.03 Gb of nonredundant repetitive sequences, accounting for 45.81% of the genome.

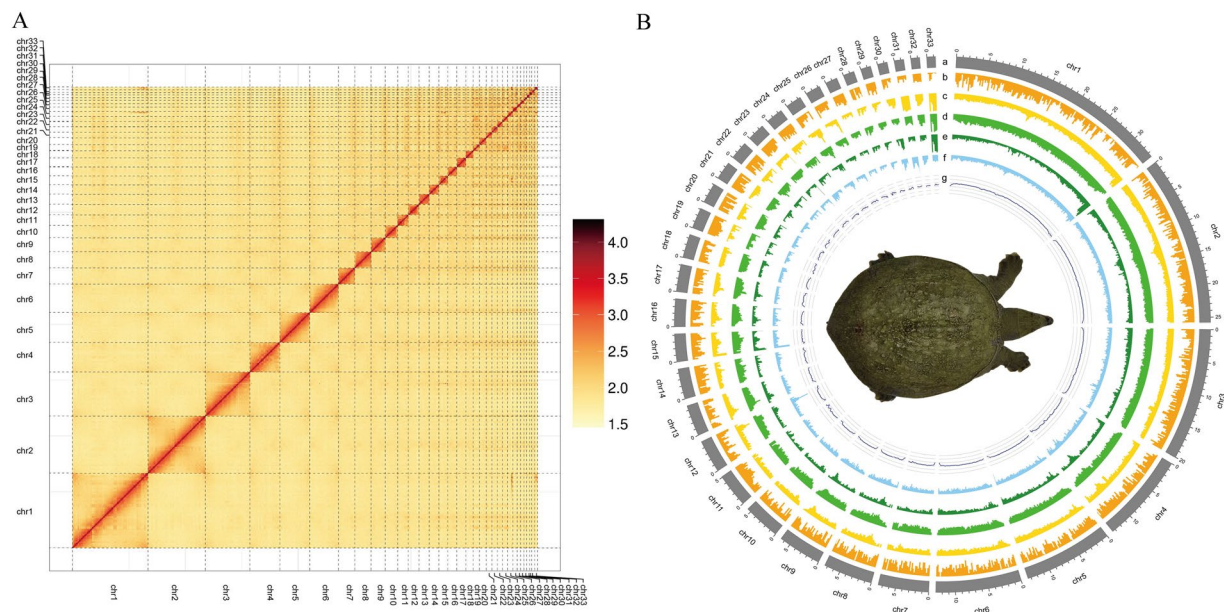


Fig. 1 Genome-wide chromosomal heatmap (A) and circos plot of genome (B). The rings from inside to outside indicate (a) pseudochromosome length of the genome, (b) gene density, (c) total transposable elements (TE) density, (d) DNA-TE density, (e) long interspersed nuclear element (LINE) density, (f) long terminal repeats (LTR) density, and (g) GC% density.

Sequences ID	Sequences Length (bp)	Sequences ID	Sequences Length (bp)
Chr1	336,740,722	Chr18	35,890,940
Chr2	257,550,244	Chr19	31,406,300
Chr3	200,317,163	Chr20	30,761,100
Chr4	134,762,063	Chr21	27,310,165
Chr5	134,669,767	Chr22	27,219,100
Chr6	129,579,767	Chr23	26,284,929
Chr7	76,334,767	Chr24	25,597,082
Chr8	74,183,000	Chr25	23,846,178
Chr9	65,814,100	Chr26	22,377,663
Chr10	55,959,633	Chr27	18,020,597
Chr11	51,079,416	Chr28	16,673,000
Chr12	49,691,045	Chr29	15,341,438
Chr13	47,044,923	Chr30	14,764,038
Chr14	46,106,945	Chr31	14,320,391
Chr15	42,062,300	Chr32	13,258,138
Chr16	41,798,940	Chr33	13,042,060
Chr17	41,338,924	—	—
Total	2,243,870,947	Percentage	95.42%

Table 1. Statistics of assembled chromosomes sequence length.

For noncoding RNA (ncRNA) annotation, rRNA and tRNA prediction was conducted using RNAmmer v 1.2⁴⁴ and tRNAscan v 1.3⁴⁵, respectively. Furthermore, other ncRNAs were detected using Rfam database⁴⁶. Six types of ncRNAs, including 24 lncRNAs, 837 miRNAs, 2958 rRNAs, 721 snRNAs, 10 ribozymes, and 7394 tRNAs, were identified from the *P. sinensis* genome (Table 4).

Gene prediction and functional annotation. The gene structures were predicted according to three approaches, including *de novo*-based, homology-based, and RNA-seq-based prediction, were used to identify gene structure. For *de novo*-based prediction, gene prediction was performed using AUGUSTUS v 3.4.0⁴⁷, GlimmerHMM v 3.0.4⁴⁸, Genscan v 3.1⁴⁹, GeneID v 1.4⁵⁰, and SNAP (version 2006-07-28)⁵¹ with default parameters. The protein sequences of *Alligator sinensis*, *Chelonia mydas*, *Chrysemys picta bellii*, *Deinagkistrodon acutus*, *Gallus gallus*, *Gekko japonicus*, and *P. sinensis* (previously published)¹³ were downloaded from Ensembl⁵². Homology-based predictions were performed with protein sequences from these reference species. For the

Sample ID	Length		Number	
	Contig** (bp)	Scaffold (bp)	Contig**	Scaffold
Total	2,243,866,247	2,243,870,947	322	275
Max	196,392,900	336,740,722	—	—
Number ≥ 2000	—	—	317	270
N50	107,607,917	129,579,767	8	6
N60	64,709,364	65,814,100	11	9
N70	46,106,945	47,044,923	16	13
N80	22,377,663	35,890,940	23	18
N90	12,489,110	22,377,663	36	26

Table 2. Statistics of *P. sinensis* genome assembly.

Type	Genome Assembly		Protein-coding gene models	
	Number	Rate (%)	Number	Rate (%)
Complete BUSCOs (C)	3260	97.2	3226	96.2
Complete and single-copy BUSCOs (S)	3216	95.9	3199	95.4
Complete and duplicated BUSCOs (D)	44	1.3	27	0.8
Fragmented BUSCOs (F)	27	0.8	51	1.5
Missing BUSCOs (M)	67	2.0	77	2.3

Table 3. BUSCO evaluation of *P. sinensis* genome.

Repeat type	<i>De novo</i> + Repbase Length (bp)	Proportion in Genome (%)
DNA	307,542,416	13.71
LINE	347,047,979	15.47
SINE	15,848,420	0.71
LTR	133,306,553	5.94
Satellite	2,428,989	0.11
Simple_repeat	604,564	0.03
Unknown	221,119,040	9.85
Total	1,027,897,961	45.81
ncRNA type	Copy	Proportion in Genome (%)
lncRNA	24	0.00%
miRNA	837	0.00%
rRNA	2958	0.26%
snRNA	721	0.00%
ribozyme	10	0.00%
tRNA	7394	0.02%

Table 4. Classification of repetitive sequences and ncRNAs.

RNA-seq-based method, the full-length transcriptome sequences generated from PacBio sequencing were aligned to the genome using the TopHat v 2.1.1⁵³, and gene structure was predicted using Cufinks v 2.2.1⁵⁴. All the gene models were merged, and redundancy was removed using MAKER2⁵⁵. Overall, a total of 21,532 protein-coding genes were predicted with an average transcript length of 40,287.42 bp, average cds length of 1597.32 bp, average intron length of 167.95 bp, average exon length of 4546.19 bp, and average exons per gene of 9.51 (Table 5).

For functional annotation, the Diamond v 2.0.6⁵⁶ was used to align all protein-coding genes to the non-redundant protein (NR) and Swissprot databases with an E-value threshold of $1e-5$. The annotation of Gene Ontology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways was performed by Blast2GO⁵⁷. The protein motifs and domains were identified using the Pfam⁵⁸.

A total of 21,149 genes (98.22% of the predicted protein-coding genes) were annotated using the above databases, and approximately 89.59%, 82.51%, 94.79%, 84.01%, and 65.34% were annotated in Swissprot, Pfam, Nr, KEGG, and GO, respectively (Table 5). A total of 12,880 genes were commonly annotated by these databases (Fig. 2).

Ethics statement. This study was approved by the the Institutional Animal Care and Use Committee (IACUC) of the Zhejiang Institute of Freshwater Fisheries. All the methods used in this study were conducted following approved guidelines.

Gene structure annotation	
Number of protein-coding gene	21,532
Average transcript length (bp)	40,287.42
Average exons per gene	9.51
Average exon length (bp)	167.95
Average CDS length (bp)	1597.32
Average intron length (bp)	4546.19
Gene function annotation	
Number (Percent)	
Swissprot	19,290 (89.59%)
Pfam	17,766 (82.51%)
Nr	20,411 (94.79%)
KEGG	18,090 (84.01%)
GO	14,069 (65.34%)
In_all_DB	12,880 (59.80%)
Annotated	21,149 (98.22%)
Total	21,532 (100%)

Table 5. Statistics of gene structure and functional annotation of *P. sinensis* genome.

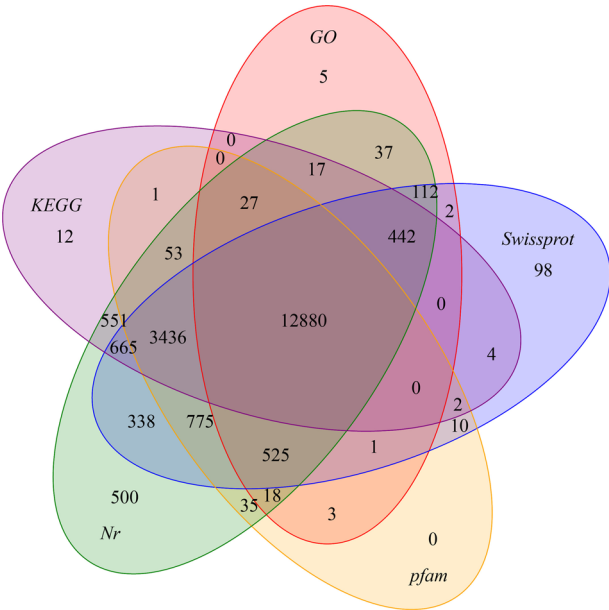


Fig. 2 Venn diagram of the number of genes from *P. sinensis* genome functional classification using multiple public databases.

Data Records

All the raw sequencing data utilized in this study were submitted to the National Center for Biotechnology Information (NCBI) SRA (Sequence Read Archive) database under BioProject accession number PRJNA1149904. the Illumina WGS data, PacBio HiFi data, Iso-Seq and Hi-C data was deposited with the accession number SRR30305005⁵⁹, SRR30305004⁶⁰, SRR30323617⁶¹ and SRR30305006⁶², respectively. The RNA-seq data have been were archived under the accession numbers SRR30304998⁶³, SRR30304999⁶⁴, SRR30305000⁶⁵, SRR30305001⁶⁶, SRR30305002⁶⁷, SRR30305003⁶⁸ in the kidney, spleen, lung, muscle, liver and heart tissues, respectively. The genome assembly has also been deposited at NCBI with the accession number GCA_049634645.1⁶⁹. The genome annotation have been deposited at the Figshare⁷⁰.

Technical Validation

To verify the integrity and accuracy of the genome assembly, the BUSCO v5.4.6 assessment was conducted with the vertebrata_odb10 database, the final genome assembly demonstrated a BUSCO completeness of 97.2%, with 95.9% single-copy BUSCOs, 1.3% duplicated BUSCOs, 0.8% fragmented BUSCOs, and 2.0% missing BUSCOs (Table 3). Furthermore, the PacBio Hifi reads were mapped to the genome using BWA and counted for mapping ratio. As a result, the mapping ratio of the assembly were 98.43%, and the genome coverage of the assembly were 99.66%. In addition, a total of 21,532 nonredundant protein-coding genes were successfully produced by combining *de novo*-based, homology-based, and RNA-seq-based prediction. A total of 21,149 genes were

successfully functionally annotated. Therefore, the high mapping ratio, genome coverage, recognition rate of single-copy orthologues and gene number indicated the high-quality of *P. sinensis* genome.

Code availability

All data processing commands and pipelines were carried out according to the instructions and guidelines of the corresponding bioinformatics software. This study does not involve specific code or script.

Received: 8 April 2025; Accepted: 11 August 2025;

Published online: 29 September 2025

References

- Liang, Y. *et al.* Establishment and population genetic analysis of SNP fingerprinting of Chinese soft-shelled turtle (*Pelodiscus sinensis*). *Aquacult Rep* **38**, 102340 (2024).
- Bu, X., Liu, L. & Nie, L. Genetic diversity and population differentiation of the Chinese soft-shelled turtle (*Pelodiscus sinensis*) in three geographical populations. *Biochem Syst Ecol* **54**, 279–284 (2014).
- Zhang, H. Q. *et al.* Differentiation of four strains of Chinese soft-shelled turtle (*Pelodiscus sinensis*) based on high-resolution melting analysis of single nucleotide polymorphism sites in mitochondrial DNA. *Genet Mol Res* **14**, 13144–13150 (2015).
- Chen, J. *et al.* Complete Mitochondrial Genomes of Four *Pelodiscus sinensis* Strains and Comparison with Other Trionychidae Species. *Biology (Basel)* **12**, 406 (2023).
- FAO Fisheries and Aquaculture, FAO Yearbook Fishery and Aquaculture Statistics 2024, Food and Agriculture Organization of the United Nations, Rome (2024).
- He, Y. *et al.* Twenty microsatellite loci from Chinese soft-shelled Turtles *Trionyx sinensis*, a vulnerable species on the IUCN Red List. *Conservation Genet Resour* **10**, 13–15 (2018).
- IUCN Red List. Available online: <https://www.iucnredlist.org/species/39620/97401140>.
- Qi, M. *et al.* Investigation of Plasticity in Morphology, Organ Traits and Nutritional Composition in Chinese Soft-Shelled Turtle (*Pelodiscus sinensis*) Under Different Culturing Modes. *Fishes* **10**, 89 (2025).
- Li, H. *et al.* Phylogenetic relationships and divergence dates of softshell turtles (Testudines: Trionychidae) inferred from complete mitochondrial genomes. *J Evol Biol* **30**, 1011–1023 (2017).
- Hong, X. *et al.* A chromosome-level genome assembly of the Asian giant softshell turtle *Pelochelys cantorii*. *Sci Data* **10**, 754 (2023).
- Grueber, C. E. & Sunnucks, P. Using genomics to fight extinction. *Science* **376**, 574–575 (2022).
- Supple, M. A. & Shapiro, B. Conservation of biodiversity in the genomics era. *Genome Biol* **19**, 131 (2018).
- Wang, Z. *et al.* The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle-specific body plan. *Nat Genet* **45**, 701–706 (2013).
- Liu, X. *et al.* Chromosome-level genome assembly of Asian yellow pond turtle (*Mauremys mutica*) with temperature-dependent sex determination system. *Sci Rep* **12**, 7905 (2022).
- Liu, J. *et al.* Chromosome-level genome assembly of the Chinese three-keeled pond turtle (*Mauremys reevesii*) provides insights into freshwater adaptation. *Mol Ecol Resour* **22**, 1596–1605 (2022).
- Ren, Y. *et al.* Genomic insights into the evolution of the critically endangered soft-shelled turtle *Rafetus swinhoei*. *Mol Ecol Resour* **22**, 1972–1985 (2022).
- Tollis, M. *et al.* The Agassiz's desert tortoise genome provides a resource for the conservation of a threatened species. *PLoS One* **12**, e0177708 (2017).
- Brian Simison, W., Parham, J. F., Papenfuss, T. J., Lam, A. W. & Henderson, J. B. An Annotated Chromosome-Level Reference Genome of the Red-Eared Slider Turtle (*Trachemys scripta elegans*). *Genome Biol Evol* **12**, 456–462 (2020).
- Cao, D., Wang, M., Ge, Y. & Gong, S. Draft genome of the big-headed turtle *Platysternon megacephalum*. *Sci Data* **6**, 60 (2019).
- Shaffer, H. B. *et al.* The western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage. *Genome Biol* **14**, R28 (2013).
- Quesada, V. *et al.* Giant tortoise genomes provide insights into longevity and age-related disease. *Nat Ecol Evol* **3**, 87–95 (2019).
- Liu, X. *et al.* Chromosome-Level Analysis of the *Pelochelys cantorii* Genome Provides Insights to Its Immunity, Growth and Longevity. *Biology (Basel)* **12**, 939 (2023).
- Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
- Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* **37**, 1155–1162 (2019).
- Belton, J. M. *et al.* Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268–276 (2012).
- van Berkum, N. L. *et al.* Hi-C: a method to study the three-dimensional architecture of genomes. *J Vis Exp*, 1869 (2010).
- Wang, H. *et al.* Estimation of genome size using *k*-mer frequencies from corrected long reads. arXiv: Genomics (2020).
- Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun* **11**, 1432 (2020).
- Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**, 170–175 (2021).
- Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* **19**, 460 (2018).
- Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**, 95–98 (2016).
- Dudchenko, O. *et al.* De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
- Robinson, J. T. *et al.* Juicebox.js Provides a Cloud-Based Visualization System for Hi-C Data. *Cell Syst* **6**, 256–258.e1 (2018).
- Hiroiyuki, S. & Hidetoshi, O. Karyotype of the Chinese soft-shelled turtle, *Pelodiscus sinensis*, from Japan and Taiwan, with chromosomal data for *Dogania subplana*. *Curr Herpetol* **20**, 19–25 (2001).
- Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res* **19**, 1639–1645 (2009).
- Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
- Seppy, M., Manni, M. & Zdobnov, E. M. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods Mol Biol* **1962**, 227–245 (2019).
- Bai, Y. *et al.* Chromosome-Level Assembly of the Southern Rock Bream (*Oplegnathus fasciatus*) Genome Using PacBio and Hi-C Technologies. *Front Genet* **12**, 811798 (2021).
- Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci USA* **117**, 9451–9457 (2020).
- Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res* **35**, W265–W268 (2007).
- Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* **6**, 11 (2015).

42. Price, A. L., Jones, N. C. & Pevzner, P. A. *De novo* identification of repeat families in large genomes. *Bioinformatics* **21**, i351–i358 (2005).
43. Behboudi, R., Nouri-Baygi, M. & Naghibzadeh, M. RPTK: A rapid perfect tandem repeat finder tool for DNA sequences. *Biosystems* **226**, 104869 (2023).
44. Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* **35**, 3100–3108 (2007).
45. Chan, P. P., Lin, B. Y., Mak, A. J. & Lowe, T. M. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res* **49**, 9077–9096 (2021).
46. Kalvari, I. *et al.* Non-Coding RNA Analysis Using the Rfam Database. *Curr Protoc Bioinformatics* **62**, e51 (2018).
47. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19**, ii215–ii225 (2003).
48. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
49. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268**, 78–94 (1997).
50. Alioto, T., Blanco, E., Parra, G. & Guigó, R. Using geneid to Identify Genes. *Curr Protoc Bioinformatics* **64**, e56 (2018).
51. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
52. Harrison, P. W. *et al.* Ensembl 2024. *Nucleic Acids Res* **52**, D891–D899 (2024).
53. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
54. Ghosh, S. & Chan, C. K. Analysis of RNA-Seq Data Using TopHat and Cufflinks. *Methods Mol Biol* **1374**, 339–361 (2016).
55. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
56. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**, 59–60 (2015).
57. Conesa, A. & Götz, S. Blast2GO: A comprehensive suite for functional analysis in plant genomics. *Int J Plant Genomics* **2008**, 619832 (2008).
58. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res* **49**, D412–D419 (2021).
59. NCBI Sequence Read Archive. <https://identifiers.org/ncbi/insdc.sra:SRR30305005> (2025).
60. NCBI Sequence Read Archive. <https://identifiers.org/ncbi/insdc.sra:SRR30305004> (2025).
61. NCBI Sequence Read Archive. <https://identifiers.org/ncbi/insdc.sra:SRR30323617> (2025).
62. NCBI Sequence Read Archive. <https://identifiers.org/ncbi/insdc.sra:SRR30305006> (2025).
63. NCBI Sequence Read Archive. <https://identifiers.org/ncbi/insdc.sra:SRR30304998> (2025).
64. NCBI Sequence Read Archive. <https://identifiers.org/ncbi/insdc.sra:SRR30304999> (2025).
65. NCBI Sequence Read Archive. <https://identifiers.org/ncbi/insdc.sra:SRR30305000> (2025).
66. NCBI Sequence Read Archive. <https://identifiers.org/ncbi/insdc.sra:SRR30305001> (2025).
67. NCBI Sequence Read Archive. <https://identifiers.org/ncbi/insdc.sra:SRR30305002> (2025).
68. NCBI Sequence Read Archive. <https://identifiers.org/ncbi/insdc.sra:SRR30305003> (2025).
69. NCBI Assembly https://identifiers.org/ncbi/insdc.gca:GCA_049634645.1 (2025).
70. Chen, J. Genome annotation of Chinese soft-shelled turtle (*Pelodiscus sinensis*). *Figshare* <https://doi.org/10.6084/m9.figshare.28715903> (2025).

Acknowledgements

This work was supported by the Key Scientific and Technological Grant of Zhejiang for Breeding New Agricultural Varieties (No: 2021C02069-8), Zhejiang Province Agricultural Major Technology Collaborative Promotion Plan Project (No: 2024ZDXT16).

Author contributions

H.Z. and J.C. conceived and designed the study. J.C. and J.B. collected the samples. X.Y., L.H., X.B. and X.P. performed the data analysis. J.C. wrote the manuscript. H.Z., J.Y. and J.C. revised the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to H.Z.

Reprints and permissions information is available at www.nature.com/reprints.

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



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/>.

© The Author(s) 2025