scientific data

Check for updates

OPEN

DATA DESCRIPTOR

An improved chromosomal-scale genome assembly of the Tanaka's snailfish (*Liparis tanakae*)

Yunlong Chen[1,2,3], Yue Jin[1,2,3], Chengcheng Su[1,2], Fayang Zhang[1] & Xiujuan Shan[1,2,3] ✉

As one of the top predators in the Yellow Sea, the Tanaka's snailfish (*Liparis tanakae*) plays an important ecological role in maintaining the structure and function of the ecosystem. This species also has fast and strong adaptability to external pressures such as climate change and fishing activities. To facilitate further molecular evolution researches of *L. tanakae*, we generated a chromosome-scale genome assembly in this study. The final assembly yielded 574.44 Mb in total length, with a scaffold N50 of 24.64 Mb, and anchored 97.87% of the sequences into 24 pseudo-chromosomes. Our assembly was 20.18 Mb longer than the reference genome (Tanakav1) in total length, with higher scaffold N50 and fewer scaffolds. The BUSCO score of 97.3% and Merqury quality value of 36.98 revealed high completeness and accuracy of our assembly. The genome contained 20,933 predicted protein-coding genes and 28.28% of the assembly was annotated as repetitive sequences. This study significantly advances the genomic resources for *L. tanakae* and facilitates future adaptation and evolution researches of this species.

Background & Summary

External issues such as fishing activities, climate change, habitat degradation and pollution are largely threatening marine biodiversity and ecosystem stability[1], leading to a widespread decline in global fishery resources[2,3]. As one of the most active fishing areas in China, the Yellow Sea is currently experiencing significant fisheries resource declines due to anthropogenic activities[4]. As a result, the Yellow Sea ecosystem has showed rapid responses to these above-mentioned pressures, particularly a continual change in fish community structure with frequent replacement of dominant species[5]. However, the Tanaka's snailfish (*Liparis tanakae*) has been a dominant species since the 1980s, flourishing for more than 40 years with a relatively stable population size[4].

In addition to its long-term dominance, the Tanaka's snailfish is also one of the top predators, playing an important ecological role in regulating the biomass of other species through a top-down effect[6]. Therefore, the Tanaka's snailfish is of high ecological importance, contributing to maintain the stability of the Yellow Sea ecosystem. Besides, the Tanaka's snailfish belongs to the family Liparidae, which contains species survive in extreme hadal environments[7–9]. Considering the Tanaka's snailfish inhabits muddy bottom regions at depths of 50–90 m[5], comparative analyses between the Tanaka's snailfish and hadal snailfish species can help to decipher the evolutionary mechanisms of vertebrates adapt to the hadal environments[9], which is of great importance in exploring the survival strategies of organisms in extreme environments.

Previous studies generally focused on ecological aspects of the Tanaka's snailfish[4–6,10–12], yet limited genetic and genomic resources have largely constrained evolutionary investigations of this species. For example, molecular mechanisms underlying the long-term dominance of the Tanaka's snailfish are still unknown, mainly due to the limited availability of genomic data. Besides, genetic information such as the degree of genetic diversity, evolutionary history and population genetic differentiation, which could provide valuable reference information for fishery resource management and conservation of the Tanaka's snailfish, are poorly understood. The development of sequencing techniques and genome-scale analytical approaches have facilitated genomics studies of marine fish species[13,14], including phylogenomics[15], population genomics[16], evolutionary genomics[17],

[1]State Key Laboratory of Mariculture Biobreeding and Sustainable Goods, Yellow Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Qingdao, 266071, China. [2]Laboratory for Marine Fisheries Science and Food Production Processes, Qingdao Marine Science and Technology Center, Qingdao, 266237, China. [3]Shandong Changdao National Observation and Research Station for Fisheries Resources, Yantai, 265800, China. ✉e-mail: shanxj@ysfri.ac.cn

Sequencing technology	Library size (bp)	Raw data (Gb)	Clean data (Gb)	Coverage (X)	Mean subread length (bp)
Illumina	350	51.08	43.00	75.44	144
PacBio	20,000	26.95	—	47.28	16141
Hi-C	350	53.35	49.00	85.96	144
RNA-seq	350	22.69	21.75	38.16	144

Table 1. Sequencing data for the Tanaka's snailfish genome assembly.

	Total length (bp)	Number of contigs	Contig N50 (bp)	Number of scaffolds	Scaffold N50 (bp)
PacBio sequencing	574,967,233	1,626	1,346,787	—	—
Hi-C sequencing	574,442,414	—	—	126	24,637,528
Genome annotation					
Protein-coding gene	20,933 (20,376 annotated, 97.34%)				
Repeat ratio	28.28%				
GC content	41.26%				
Genome quality assessment					
Illumina reads mapping rate	98.10%				
Illumina reads coverage	99.66%				
Merqury QV-value	36.98				
BUSCO evaluation	n = 3,354				
Complete BUSCOs	3,210 (95.7%)				
Complete and single-copy BUSCOs	3,116 (92.9%)				
Complete and duplicated BUSCOs	94 (2.8%)				
Fragmented BUSCOs	54 (1.6%)				
Missing BUSCOs	90 (2.7%)				

Table 2. Assembly and annotation statistics of the Tanaka's snailfish genome.

conservation genomics[18], and among others. Till now, a total of two genome assemblies of the Tanaka's snailfish have been deposited in the NCBI Genome database: one chromosome-level assembly (GenBank accession no. GCA_036178185.1) and one scaffold-level assembly (GCA_006348945.1). Although the chromosome-level assembly was chosen as the reference genome, it has a total of 926 scaffolds, including 24 chromosomes and 902 unplaced scaffolds. Besides, the reference genome sequence was only assembled using Oxford Nanopore sequencing data, lacking processes such as gap closing and genome polishing based on short-reads sequencing data, to some extent impacting the continuity of the assembly.

In this study, we assembled a chromosome-scale genome sequence of the Tanaka's snailfish using Illumina short reads, PacBio HiFi long reads and Hi-C data (Table 1). The initial genome assembly had a total length of 574.97 Mb with 1,626 contigs and a contig N50 of 1.35 Mb (Table 2). After Hi-C scaffolding approach, 97.87% of the initial assembled sequences were anchored to 24 pseudo-chromosomes (Fig. 1), and the total length of the final genome assembly was 574.44 Mb, with 126 scaffolds and scaffold N50 of 24.64 Mb (Table 2). Our assembly was 20.18 Mb longer than the NCBI reference genome in total length, with higher scaffold N50 (24.64 vs. 23.04 Mb), fewer scaffolds (126 vs. 926) and higher chromosome anchoring rate (Table 3), showing relatively high assembly quality. Higher assembly completeness, continuity and integrity were also observed when comparing to the scaffold-level assembly GCA_006348945.1 (Table 3). In our assembled sequence, a total of 162.47 Mb of repetitive sequences were annotated, representing 28.28% of the genome assembly (Table 2). The repetitive sequences (Table 4) were dominated by DNA transposons (46.32 Mb, 8.06%), long interspersed elements (LINEs, 28.93 Mb, 5.04%) and long terminal repeats (LTRs, 12.03 Mb, 2.09%). In addition, combining *ab initio*, homology-based and RNA-seq assisted gene prediction approaches, a total of 20,933 protein-coding genes were predicted, among which 20,376 (97.11%) were annotated (Table 2, Fig. 2). A total of 46,587 non-coding RNA (ncRNA) genes were predicted, including 1,583 miRNAs, 32,466 tRNAs, 10,955 rRNAs and 1,583 snRNAs (Table 5). The assembled genome sequence and associated annotation information provide valuable resources for elucidating the genetic adaptation and underlying molecular basis of the long-term dominance of Tanaka's snailfish. These genomic data can be also used in future comparative genomics studies to investigate genomic evolution and phylogeny of snailfishes.

Methods

Sample collection and sequencing. An adult female Tanaka's snailfish individual was sampled from the Yellow Sea (123°10′E, 38°33′N) in May 2023. The muscle tissue below the dorsal fin was taken and stored in the liquid nitrogen until DNA extraction. Genomic DNA was isolated using the cetyltrimethylammonium bromide (CTAB) method. High-quality DNA was used for library preparation and high-throughput sequencing.

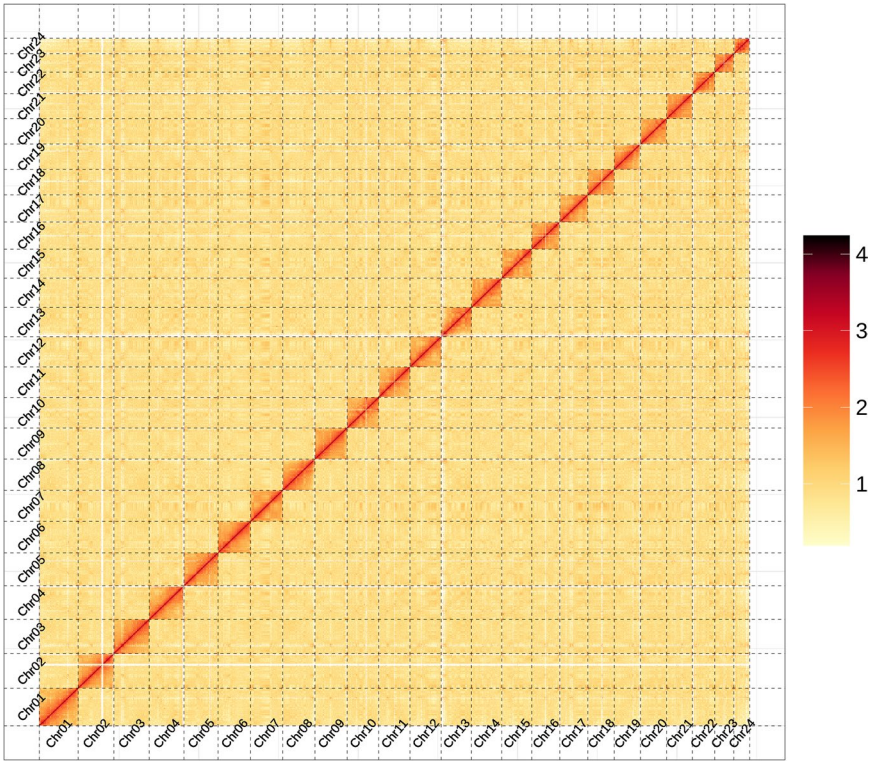

Fig. 1 The Hi-C contact map of the Tanaka's snailfish genome assembly in this study. chr 1–24 represented for the 24 pseudo-chromosomes. The color bar showed the contact density from white (low) to black (high).

	This study	GCA_036178185.1	GCA_006348945.1
Total length (bp)	574,442,414	554,262,514	498,979,456
Number of chromosome	24	24	—
Number of scaffold	126	926	27,878
Unplaced scaffold	102	902	—
Chromosome anchoring rate	97.87%	95.47%	—
Scaffold N50 (bp)	24,637,528	23,043,945	375,216
Number of contig	1,626	1,335	97,972
Contig N50 (bp)	1,346,787	5,908,986	9,903
GC percent	41.26%	43.5%	43.5%
Assembly level	Chromosome	Chromosome	Scaffold
Protein-coding gene	20,933 (20,376 annotated)	—	68,289 (68,269 annotated)
Sequencing technology	Illumina, PacBio HiFi, Hi-C	Oxford Nanopore	Illumina

Table 3. Comparison of assembly statistics of three Tanaka's snailfish genome sequences.

	Repeat size (bp)	Percentage of genome (%)
DNA	46,319,007	8.06
LINE	28,932,208	5.04
SINE	4,289,844	0.75
LTR	12,030,884	2.09
Unknown	64,642,557	11.25
Other	6,258,850	1.09
Total	162,473,350	28.28

Table 4. Statistics of repetitive sequences in the Tanaka's snailfish genome assembly.

Illumina short-insert (350 bp) libraries were prepared according to the protocol and paired-end (PE150) sequenced on the Illumina Novaseq 6000 platform (Illumina, Inc., San Diego, CA, USA). HiFi long-read sequencing was performed using the PacBio Sequel II sequencer (Pacific Biosciences, Menlo Park, CA, USA).

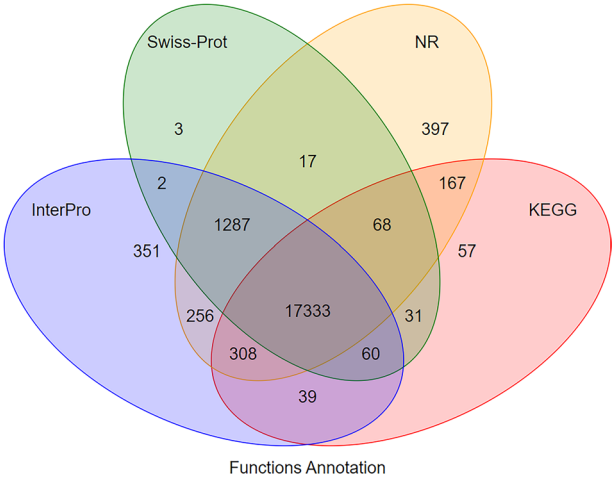

Fig. 2 Venn diagram of functional annotation of the Tanaka's snailfish genome assembly in this study.

Type		Number	Average length (bp)	Total length (bp)	Proportion in Genome (%)
miRNA		1,583	153.21	242,540	0.04
tRNA		32,466	75.57	2,453,743	0.42
rRNA	Total	10,955	131.96	1,445,699	0.25
	18S	233	727.30	169,463	0.02
	28S	748	421.79	315,506	0.05
	5.8S	94	152.81	14,365	0.002
	5S	9,880	95.78	946,365	0.16
snRNA	Total	1,583	153.21	242,540	0.04
	CD-box	361	150.42	54,303	0.009
	HACA-box	120	159.27	19,113	0.003
	splicing	1,008	150.89	152,100	0.02
	other	85	194.24	16,511	0.002

Table 5. Classification of ncRNA genes in the Tanaka's snailfish genome assembly in this study.

For Hi-C sequencing, fresh muscle was fixed with formaldehyde in a concentration of 1% and the fixation was terminated using 0.2 M glycine. A Hi-C library was prepared following the Hi-C library protocol[19] and then sequenced using an Illumina Novaseq 6000 sequencing platform. We also constructed four RNA-seq libraries to facilitate prediction of protein-coding genes. The RNA-seq libraries were then sequenced on an Illumina sequencing platform.

Genome assembly. A total of 26.95 Gb PacBio HiFi long-read data (Table 1) were used for *de novo* genome assembly using Hifiasm[20] with default parameters. Genome polishing was performed using BWA v0.7.10[21] and Pilon v1.23[22] with Illumina short reads (clean data 43.00 Gb, Table 1). These sequencing data resulted in a 574.97 Mb assembly with 1,626 contigs and a contig N50 of 1.35 Mb (Table 2). The draft genome contigs were then anchored and oriented into a chromosomal-scale assembly using the Hi-C data. A total of 49.00 Gb clean data (Table 1) were aligned to the draft genome assembly using BWA. Duplication removal, sorting, and quality control were performed using HiC-Pro v2.8.0[23]. Only uniquely mapped valid read pairs were used for further analysis. LAchesis[24] was then used to cluster, order, and orient the contigs into chromosomal-scale assembly. Finally, 97.87% of the initial assembled sequences were anchored to 24 pseudo-chromosomes (Fig. 1), and the total length of the genome assembly was 574.44 Mb, with 126 scaffolds and scaffold N50 of 24.64 Mb (Table 2).

Repetitive sequence annotation. A combined strategy based on homology alignment and *de novo* search was applied in our repeat annotation pipeline. A *de novo* repetitive elements database was built by LTR_FINDER[25], RepeatScout[26], RepeatModeler (www.repeatmasker.org/RepeatModeler.html) with default parameters. Tandem repeats were also *ab initio* extracted using TRF v4.09[27]. Then all repeat sequences with lengths >100 bp and gap 'N' less than 5% constituted the raw transposable element (TE) library. The homolog-based prediction commonly searched against Repbase[28] database employing RepeatMasker v3.3.0[29] software and its in-house scripts RepeatProteinMask (v3.2.2) with default parameters. The combination of Repbase and our *de novo* TE library was processed by uclust[30] to yield a non-redundant library and RepeatMasker was used to identify DNA-level repeat. The results of repetitive sequence annotation were listed in Table 4.

Protein-coding gene prediction and annotation. We employed *ab initio*, homology-based and RNA-seq assisted prediction to detect the protein-coding genes. For homology-based prediction, protein sequences of *Gasterosteus aculeatus, Oryzias sinensis*, *Gadus morhua*, *Danio rerio and Takifugu rubripes* were downloaded from Ensembl database[31]. The protein sequences were aligned against the genome assembly using TBLASTN v2.2.26[32] (E-value ≤ 1e-5), and then matching proteins were aligned to the homologous genome sequences for accurate spliced alignments with GeneWise v2.4.1[33]. The *ab initio* prediction was performed using Augustus v3.2.3[34], GeneID v1.4[35], GeneScan v1.0[36], GlimmerHMM v3.04[37], and SNAP v2013-11-29[38] based on the repeat masked genome sequences. RNA-seq data were mapped to the genome using HISAT2 v2.1.0[39]. Transcript structures were predicted using Stringtie v1.3.3[40], and candidate coding regions were predicted using TransDecoder v. 5.5.0 (https://github.com/TransDecoder/TransDecoder). Finally, genes predicted by the above three methods were merged into a non-redundant reference gene set with EvidenceModeler v1.1.1[41] with identical weights, leading to a total of 20,933 protein-coding genes (Table 2).

Protein-coding genes were annotated by aligning the gene sequences to the SwissProt, NT, NR, Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases using BLAST + v2.2.28[42] with an e-value threshold of 1e-5. InterProScan v5.31[43] was used to predict protein function based on conserved domains and motif by searching against ProDom, PRINTS, Pfam, SMRT, PANTHER and PROSITE. Ultimately, 20,376 (97.11%) predicted genes were successfully annotated (Table 2, Fig. 2).

For noncoding RNA (ncRNA) annotation, Infernal[44] (v1.1.4) was utilized based on the Rfam database (http://eggnogdb.embl.de/). Four types of ncRNA were identified from the Tanaka's snailfish genomes (46,587 genes in total), including 1,583 miRNAs, 32,466 tRNAs, 10,955 rRNAs and 1,583 snRNAs (Table 5).

Data Records

The sequencing dataset and genome assembly were deposited in public repositories. The raw sequencing data including Illumina, PacBio, Hi-C and RNA-seq data were submitted to the National Center for Biotechnology Information (NCBI) SRA database under BioProject accession number PRJNA1231580[45]. The assembled genome data have been deposited at GenBank under accession JBMEBB000000000.1[46], and the associated genomic annotation results are stored in Figshare database[47].

Technical Validation

Evaluation of the quality of genomic DNA and RNA. In our DNA extraction section, the DNA quality and concentration were measured using agarose gel electrophoresis (1%), pulse field gel electrophoresis (1%) and Qubit 3.0 (Thermo Fisher Scientific, Inc., Carlsbad, CA, USA), respectively. For RNA, the integrity and quantity was evaluated using the Agilent 2100 Bioanalyzer (Agilent, USA). Subsequently, high-quality DNA and RNA were used for library preparation and high-throughput sequencing.

Evaluation of the completeness of genome assembly. The completeness of the assembled genome sequence was evaluated using BUSCO v3.0.1[48]. The BUSCO analysis against the vertebrata_odb10 database found that 97.3% of the conserved single copy orthologue genes, including 95.7% of the complete and 1.6% fragmented genes, were found in the genome assembly (Table 2). The mapping rate of Illumina short reads from same individual were used to evaluate the quality of the initial genome assembly using BWA v0.7.10. By using a total of 43.00 Gb Illumina sequencing data from the same individual, the mapped read rate and coverage were 98.10% and 99.66%, respectively (Table 2), showing high consistency of our assembly. Additionally, using the Merqury' k-mer analysis[49], the quality value (QV) scores of our assembly based on short reads were estimated as 36.98 (Table 2) and the base accuracy rates were >99.9%, indicating high assembly accuracy.

Code availability

All software used in this study are in the public domain, with parameters being clearly described in Methods. If no detail parameters were mentioned for a software, default parameters were used as suggested by developer. No custom scripts or code were employed.

Received: 21 March 2025; Accepted: 27 May 2025;
Published online: 09 June 2025

References

1. He, Q. & Silliman, B. R. Climate change, human impacts, and coastal ecosystems in the Anthropocene. *Current Biology* **29**, R1021–R1035 (2019).
2. Wilson, J. R. *et al*. Adaptive comanagement to achieve climate-ready fisheries. *Conservation Letters* **11**, e12452 (2018).
3. Carozza, D. A., Bianchi, D. & Galbraith, E. D. Metabolic impacts of climate change on marine ecosystems: Implications for fish communities and fisheries. *Global Ecology and Biogeography* **28**, 158–169 (2019).
4. Chen, Y. *et al*. Long-term changes in the spatio-temporal distribution of snailfish *Liparis tanakae* in the Yellow Sea under fishing and environmental changes. *Frontiers in Marine Science* **9**, 1024086 (2022).
5. Chen, Y. *et al*. Changes in fish diversity and community structure in the central and southern Yellow Sea from 2003 to 2015. *Journal of Oceanology and Limnology* **36**, 805–817 (2018).
6. Lin, Q., Jin, X. & Zhang, B. Trophic interactions, ecosystem structure and function in the southern Yellow Sea. *Chinese Journal of Oceanology and Limnology* **31**, 46–58 (2013).
7. Mu, Y. *et al*. Whole genome sequencing of a snailfish from the Yap Trench (~7,000 m) clarifies the molecular mechanisms underlying adaptation to the deep sea. *PLoS Genetics* **17**, e1009530 (2021).
8. Wang, K. *et al*. Morphology and genome of a snailfish from the Mariana Trench provide insights into deep-sea adaptation. *Nature Ecology and Evolution* **3**, 823–833 (2019).
9. Xu, W. *et al*. Chromosome-level genome assembly of hadal snailfish reveals mechanisms of deep-sea adaptation in vertebrates. *Elife* **12**, RP87198 (2023).

10. Chen, Y. *et al*. Estimating seasonal habitat suitability for migratory species in the Bohai Sea and Yellow Sea: A case study of tanaka's snailfish (*Liparis tanakae*). *Acta Oceanologica Sinica* **41**, 22–30 (2022).
11. Tomiyama, T., Yamada, M. & Yoshida, T. Seasonal migration of the snailfish *Liparis tanakae* and their habitat overlap with 0-year-old Japanese flounder *Paralichthys olivaceus*. *Journal of the Marine Biological Association of the United Kingdom* **93**, 1981–1987 (2013).
12. Zhou, Z., Jin, X., Shan, X., Li, Z. & Dai, F. Seasonal variations in distribution and biological characteristics of snailfish *Liparis tanakae* in the central and southern Yellow Sea. *Acta Ecologica Sinica* **32**, 5550–5561 (2012).
13. Kelley, J. L. *et al*. The life aquatic: advances in marine vertebrate genomics. *Nature Review Genetics* **17**, 523–534 (2016).
14. Ahmad, S. F. *et al*. Fish genomics and its impact on fundamental and applied research of vertebrate biology. *Reviews in Fish Biology and Fisheries* **32**, 357–385 (2022).
15. Xu, S., Zhao, R., Cai, S., Li, P. & Han, Z. Application of genomic markers generated for ray-finned fishes in chondrichthyan Phylogenomics. *Organisms Diversity & Evolution* **23**, 1005–1012 (2023).
16. Liu, Y. F., Li, Y. L., Xing, T. F., Xue, D. X. & Liu, J. X. Genetic architecture of long-distance migration and population genomics of the endangered Japanese eel. *Iscience* **27**, 110563 (2024).
17. Zhang, K. *et al*. Genomics comparisons provide new insights into the evolution of karyotype and body patterns in Anguilliformes species. *International Journal of Biological Macromolecules* **308**, 142504 (2025).
18. Humble, E. *et al*. Comparative population genomics of manta rays has global implications for management. *Molecular Ecology* **34**, e17220 (2025).
19. Rao, S. S. *et al*. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
20. Cheng, H. *et al*. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods* **18**, 170–175 (2021).
21. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
22. Walker, B. J. *et al*. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
23. Servant, N. *et al*. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biology* **16**, 259 (2015).
24. Burton, J. N. *et al*. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature Biotechnology* **31**, 1119–1125 (2013).
25. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research* **35**, W265–268 (2007).
26. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics* **21**, i351–358 (2005).
27. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**, 573–580 (1999).
28. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6**, 11 (2015).
29. Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics* **5**, 4.10.1–4.10.14 (2004).
30. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
31. Cunningham, F. *et al*. Ensembl 2019. *Nucleic Acids Research* **47**, D745–D751 (2019).
32. Gertz, E. M. *et al*. Composition-based statistics and translated nucleotide searches: Improving the TBLASTN module of BLAST. *BMC Biology* **4**, 41 (2006).
33. Doerks, T., Copley, R. R., Schultz, J., Ponting, C. P. & Bork, P. Systematic identification of novel protein domain families associated with nuclear functions. *Genome Research* **12**, 47–56 (2002).
34. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19**, ii215–225 (2003).
35. Blanco, E., Parra, G. & Guigó, R. Using geneid to identify genes. *Current Protocols in Bioinformatics* **18**, 4.3.1–4.3.28 (2007).
36. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology* **268**, 78–94 (1997).
37. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open-source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
38. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
39. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology* **37**, 907–915 (2019).
40. Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. & Salzberg, S. L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature Protocols* **11**, 1650 (2016).
41. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology* **9**, R7 (2008).
42. McGinnis, S. & Madden, T. L. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research* **32**, W20–25 (2004).
43. Mulder, N. & Apweiler, R. InterPro and InterProScan: tools for protein sequence classification and comparison. *Methods in Molecular Biology* **396**, 59–70 (2007).
44. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
45. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP567926 (2025).
46. *NCBI GenBank* https://identifiers.org/ncbi/insdc:JBMEBB000000000 (2025).
47. Chen, Y. L. *et al*. Genome annotation of *Liparis tanakae*. *Figshare* https://doi.org/10.6084/m9.figshare.28604279 (2025).
48. Waterhouse, R. M. *et al*. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular Biology and Evolution* **35**, 543–548 (2018).
49. Rhie, A. *et al*. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology* **21**, 245 (2020).

Acknowledgements

This work was funded by the National Natural Science Foundation of China (42206104), the Special Fund of Taishan Scholar Project (tsqn202103135), and the Central Public-interest Scientific Institution Basal Research Fund, CAFS (2023TD01).

Author contributions

Yunlong Chen and Xiujuan Shan conceived the study. Yue Jin and Fayang Zhang collected the samples. Fayang Zhang extracted the genomic DNA and conducted sequencing. Chengcheng Su performed bioinformatics analysis. Yunlong Chen, Xiujuan Shan and Yue Jin wrote 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 X.S.

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.