scientific data

OPEN

DATA DESCRIPTOR

Chromosome-level Genome Assembly and Annotation of the bighead beaked sandfish (*Gonorynchus abbreviatus*)

Linlin Zhao[1], Xiaotong Zhu[1], Min Zhou[1], Ruizhi Wang[2], Na Song[2], Bailin Cong[1]✉ & Wenbo Zhu[3]✉

Gonorynchus abbreviatus **(bighead beaked sandfish), is a demersal fish inhabiting the deep waters of the Northwest Pacific for which genomic resource have been limited. Here, we assembled a chromosome-level genome using DNBSEQ, PacBio, and Hi-C sequencing technologies. The final 609.20 Mb assembly achieved a scaffold N50 of 22.33 Mb and was anchored onto 27 chromosomes, with 99.07% completeness based on BUSCO assessment. Repetitive sequences comprised 32.88% of the genome, mainly DNA transposons (12.33%). We identified 22,396 protein-coding genes, of which 21,429 (95.68%) were functionally annotated. This high-quality genome constitutes a valuable resource for investigating the evolutionary history and genetic mechanisms of *G. abbreviatus*.**

Background & Summary

The bighead beaked sandfish (*Gonorynchus abbreviatus*; NCBI Taxonomy ID: 189504) is a benthic teleost belonging to the family Gonorynchidae and the order Gonorynchiformes (Fig. 1)[1]. It is distributed in the north-western Pacific and is typically associated with sandy substrates on the continental shelf and slope. The species lacks a swim bladder and exhibits a largely demersal lifestyle[2]. Members of the genus *Gonorynchus* share a conservative external morphology and are commonly referred to as sandfishes[3]; *G. abbreviatus* can be distinguished from closely related species by possessing eight pelvic-fin rays, compared with nine in related species[2,4]. Although of limited commercial value and often encountered as by-catch, *Gonorynchus* species provide useful material for resolving phylogenetic relationships within Ostariophysi and for studying traits associated with benthic and demersal modes of life[5]. Nevertheless, little is known about their reproduction, life history and genomic architecture, and no reference genome for *G. abbreviatus* has previously been available prior to this study.

Despite the rapid increase in the number of published teleost genomes, genomic resources remain unevenly distributed across ecological and taxonomic groups[6–9], In particular, many benthic and demersal fishes inhabiting continental shelf and slope environments are still poorly represented. Improving genomic coverage of such taxa is important for advancing comparative studies of teleost evolution and ecological diversification. Mitochondrial evidence places *G. abbreviatus* close to *G. greyi* (GenBank accession AP009402), yet only a handful of diagnostic markers have been reported for *G. abbreviatus* to date[10].

Here, we reported the first chromosome-level genome assembly of *G. abbreviatus*, generated using a combination of DNBSEQ, PacBio, and Hi-C sequencing technologies. This high-quality genome provides a valuable reference for investigating the evolutionary history of Gonorynchiformes and clarifying phylogenetic relationships within Ostariophysi. Moreover, it establishes a genomic foundation for future studies of benthic and demersal teleosts, including research on comparative genomics, functional biology, and conservation.

[1]Marine Ecology Research Center, First Institute of Oceanography, Ministry of Natural Resources, 266061, Qingdao, China. [2]Key Laboratory of Mariculture, Ministry of Education, Ocean University of China, Qingdao, 266061, China. [3]Binhai Center for Natural Protected Areas and Wetland Research, Shandong Marine Resource and Environment Research Institute, Yantai, 264006, China. ✉e-mail: biolin@fio.org.cn; zhuwenbo@shandong.cn

Fig. 1 Morphological characteristics of *G. abbreviatus*.

Libraries	Clean Read Number	Clean Data (Gb)	Read Length (bp)	GC Content (%)
DNBSEQ Reads	421,306,706	62.13	147	44.77
PacBio Reads	1,599,814	31.18	19,491.9	44.11
Hi-C Reads	510,704,850	76.44	150	44.04
RNA-seq Reads	71,365,464	10.7	150	50.05

Table 1. Statistics of Sequencing Read Data.

Methods

Sample collection and sequencing. A specimen of *Gonorynchus abbreviatus* was collected from the East China Sea (26.76°N, 123.32°E) using a bottom trawl on October 22, 2021. Sampling and animal handling were approved by the Animal Care and Use Committee of the First Institute of Oceanography, Ministry of Natural Resources. Genomic DNA was extracted from muscle tissue using an SDS-based method, followed by purification with the GeneJET Genomic DNA Purification Kit (Thermo Fisher Scientific, USA). Three libraries were constructed: (i) a DNBSEQ short-read library (300–350 bp insert size), sequenced on the BGI-T7 platform, (ii) a PacBio HiFi library (10–40 kb insert size) prepared with the SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences, USA) and sequenced on the PacBio Sequel II platform, and (iii) a Hi-C library prepared from liver tissue according to Belton *et al*. with minor modifications, sequenced on the BGI-T7 platform[11]. In total, 62.13 Gb of DNBSEQ short reads, 31.18 Gb of PacBio HiFi CCS reads, and 76.44 Gb of Hi-C reads were generated (Table 1).

For transcriptome sequencing, total RNA was isolated from six tissues (muscle, brain, gonad, liver, spleen, and heart) using the GeneJET RNA Purification Kit (Thermo Fisher Scientific, USA). Libraries were prepared using the NEB RNA Library Prep Kit (NEB, USA) before sequencing on the BGI-T7 platform, yielding 10.7 Gb of clean RNA-seq data that were subsequently used for genome annotation (Table 1). All library preparation and sequencing were performed by OneMore Tech Ltd. (Wuhan, China).

Genome size estimation and assembly. Based on 62.13 Gb of clean DNBSEQ short-reads, the genome size, heterozygosity and repetitive ratio were estimated using the k-mer analysis. Briefly, 17-mers were counted from the cleaned reads, and the resulting k-mer frequency distribution was analysed using GCE (v1.0.0; parameters: -c 95 -H 1) to infer genome characteristics[12]. GCE estimates genome size primarily as the total number of k-mers divided by the main peak k-mer depth by modelling the k-mer depth distribution while accounting for sequencing errors, heterozygous k-mers, and repetitive sequences. In total, 55,384,981,584 k-mers with a depth of 95 was obtained. In addition, the genome size of *G. abbreviatus* was approximately 554.81 Mb, with a heterozygosity of 0.93% and proportion of repeat sequences at 30.59% (Fig. 2).

In the initial genome assembly, HiFiasm (v0.19.6) method was used for *de novo* to assemble the genome using the HiFi reads from PacBio[13]. This preliminary assembly yielded a genome size of 688.09 Mb (Table 2). Subsequently, the redundant sequences were removed with Purge_Haplotigs (v1.0.4; parameter: -a 70 -j 80 -d 200)[14]. Based on PacBio sequencing data, a 639.21 Mb contig-level genome assembly of *G. abbreviatus* was obtained, comprising 391 contigs with contig N50 and N90 sizes of 4.17 and 0.89 Mb, respectively (Table 2). For chromosome-scale scaffording, 76.44 Gb Hi-C reads were aligned to the contigs using Bowtie2 (v2.3.4.3) with the default parameters. In total, 118.99 million paired-end reads were uniquely mapped (Table S1), of which 45.39% was retained as valid pairs (Table S2). Subsequently, contigs were assembled into the chromosome-level scaffolds using the 3D-DNA processes (v180922; parameters: -r 0) with all valid pairs, and performed manual curation with JuiceBox (v1.11.08) following the default parameters. This procedure yielded 27 chromosome-scale scaffolds, anchoring 99.88% of the assembly; four scaffolds remained unanchored. Ultimately, we obtained a 609.20 Mb chromosome-level genome assembly of *G. abbreviatus*, with contig N50 and scaffold N50 size of 4.17 Mb and 22.33 Mb, respectively (Fig. 3; Table S3).

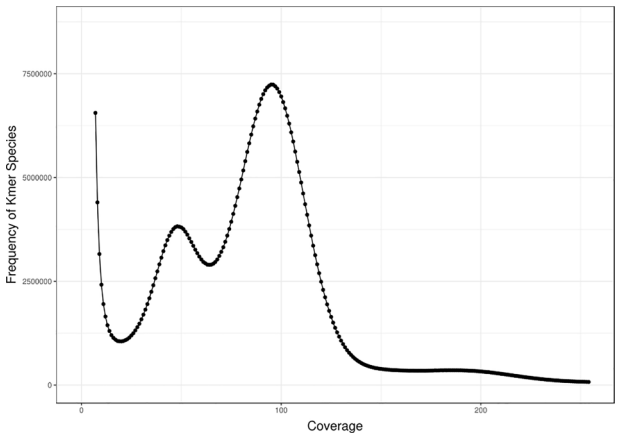

Fig. 2 K-mer (17-mer) distribution and estimation of genome size of *G. abbreviatus*.

Mode	Total Length (Mb)	Total Number	N50 (Mb)	N90 (Mb)	GC Content (%)
Hifiasm	688.09	906	3.96	0.58	45.21
Hifiasm + Purge_Halotigs	639.21	391	4.07	0.89	45.54

Table 2. Statistics for the *G. abbreviatus* genome assembly.

Repetitive sequence annotation. Repetitive sequences, including tandem repeats and interspersed repeats, in *G. abbreviatus* genome were annotated using the homology-based methods and *de novo* prediction. For homology comparison, RepeatMasker (open-4.0.9; parameters: -nolow -no_is -norna) and its built-in RepeatProteinMask (v3.2.2) were utilised to find the interspersed repeats against the RepBase database (2018-10-26)[15]. For *de novo* prediction, TRF (v4.09; parameters: 2 7 7 80 10 50 2000 -d -h) was used to identify the tandem repeats. In addition, a species-specific repetitive sequence library was constructed using the RepeatModeler (open-1.0.11) with default parameters and LTR-FINDER_parallel (v1.0.7) with default parameters. This combined repeat library was subsequently used as input for RepeatMasker (open-4.0.9) to generate a unified repeat annotation. Redundant and overlapping repeat annotations identified by different methods were resolved during the RepeatMasker-based integration, and only non-overlapping genomic regions were counted in the final statistics. As a result, a total of 200.31 Mb of repetitive sequences was identified, accounting for 32.88% of the assembled genome, with DNA transposons being the most abundant (12.33%), followed by long interspersed nuclear elements (LINEs, 5.05%) and long terminal repeats (LTRs, 5.00%) (Tables 3, 4).

Protein-coding gene annotation. First, a repeat-masked assembly was used for *de novo* gene prediction, with AUGUSTUS (v3.3.2), Genscan (v1.0) and GlimmerHMM (v3.0.4) to detect the protein-coding genes[16–18]. For homology-based evidence, proteomes from *Chanos chanos* (GCF_902362185.1), *Danio rerio* (GCF_000002035.6), *Colossoma macropomum*, (GCF_904425465.1) *and Larimichthys crocea*, (GCF_000972845.2), as well as conserved single-copy ortholog proteins from the BUSCO lineage dataset (actinopterygii_odb12), were aligned to the *G. abbreviatus* assembly with miniprot (v0.11-r234; parameters:–gff-only -O 11 -E 1 -F 23 -C 1 -B 5 -G 200000 -j 1) to define candidate loci; gene annotations from close relatives were further lifted over to *G. abbreviatus* using Liftoff (v1.6.3; parameters: -a 0.5 -s 0.5) to refine exon–intron structures[19,20]. Transcriptome-supported gene prediction was conducted by mapping RNA-seq reads to the genome using HISAT2 (v2.1.0; parameters:–dta), assembling transcripts with StringTie (v1.3.5), and identifying putative coding sequences using TransDecoder (v5.5.0). Transcript assemblies generated with Cufflinks (v2.2.1; parameters: -e 100) were used as supplementary evidence[21–23]. Evidence from *de novo* prediction, homology-based evidence (including BUSCO-derived proteins), and transcriptome-supported prediction was integrated using MAKER2 (v2.31.10; parameters: max_dna_len = 3000000, min_contig = 10000, pred_flank = 500, min_protein = 30) and HiFAP to generate a complete, non-redundant reference gene set[24]. In total, 22,396 protein-coding genes were predicted in the *G. abbreviatus* genome. Summary statistics for gene, CDS, exon and intron lengths are provided in Table 5, and gene-structure features are compared with four homologous species in Fig. 4.

Protein-coding genes were functionally annotated based on sequence similarity searches using DIAMOND (v2.0.14; parameters:–evalue 1e-05) against Swiss-Prot, TrEMBL, NR, KEGG and KOG[25–29]. Conserved domains and motifs were assigned with InterProScan (v5.61-93.0; parameters:–seqtype p–formats TSV–goterms–pathways -dp) and with HMMER3 (v3.3.1; parameters: hmmsearch -E 1e-05–domE 1e-05) against Pfam; GO terms and KEGG pathway assignments were obtained using KOBAS (v3.0; parameter: -t blastout:tab -s ko) based on sequence similarity results[30–32]. Ultimately, a total of 21,429 genes (95.68%) were successfully annotated (Table 6).

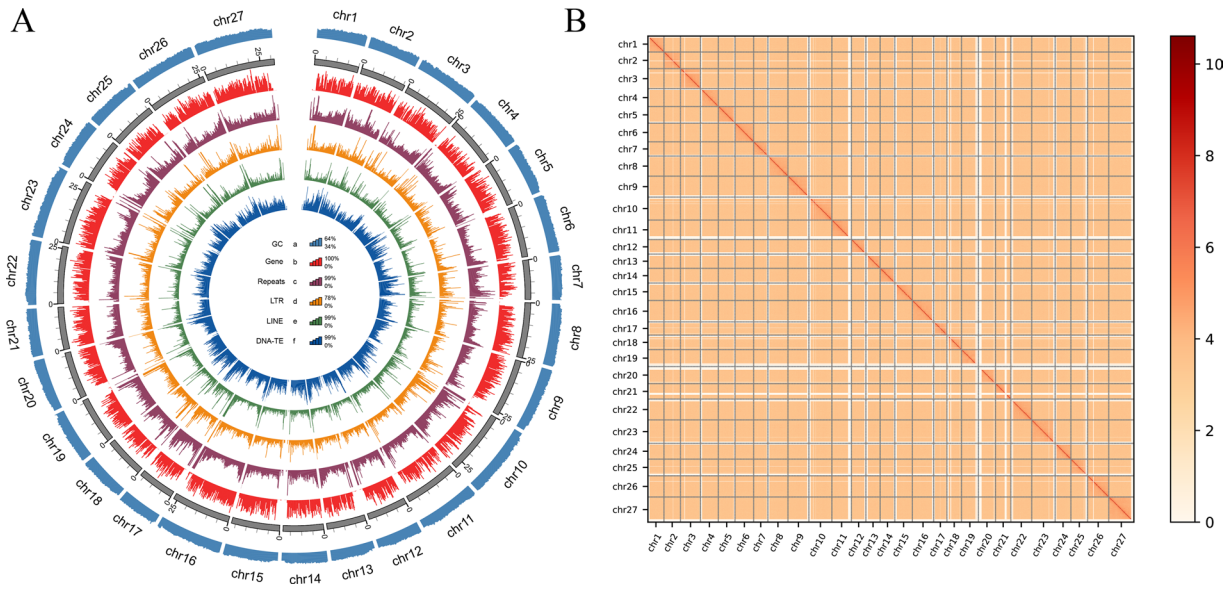

Fig. 3 (**A**) Genomic landscape of bighead beaked sandfish. The rings, from the outermost to the innermost layer, represent the chromosomes of the *G. abbreviatus* genome and GC density (a), gene density (b), total repetitive sequence density (c), LTRs density (d), LINES density (e), DNA transposons density (f). The analysis was conducted using 300 kb genomic windows. (**B**) Chromosomal Hi-C heatmap of the *G. abbreviatus* genome assembly.

	Repeat Size (Mb)	Percentage of Genome (%)
Trf	98.57	16.18
Repbase	106.83	17.54
Proteinmask	28.39	4.66
De novo	60.28	9.89
Total	200.31	32.88

Table 3. Repetitive sequences statistics for the *G. abbreviatus* genome.

	RepBase TEs	TE Proteins		*De novo*	Combined TEs		
	Length (bp)	Percentage of genome (%)	Length (bp)	Percentage of genome (%)	Length (bp)	Percentage of genome (%)	Length (bp)	Percentage of genome (%)
DNA	67,500,715	11.08	7,726,622	1.27	19,407,237	3.19	75,134,056	12.33
LINE	24,632,705	4.04	13,884,793	2.28	12,518,690	2.05	30,792,484	5.05
SINE	3,479,294	0.57	0	0.00	1,927,262	0.32	4,623,535	0.76
LTR	22,060,222	3.62	6,785,953	1.11	10,364,288	1.70	30,456,132	5.00
Satellite	6,052,913	0.99	0	0.00	390,433	0.06	6,340,527	1.04
Simple repeat	0	0.00	0	0.00	49,346	0.01	49,346	0.01
Other	2,532	0.00	0	0.00	0	0.0	2,532	0.00
Unknown	786,371	0.13	4,029	0	16,038,502	2.63	16,808,381	2.76
Total	106,873,501	17.54	28,386,633	4.66	60,276,593	9.89	145,000,031	23.80

Table 4. Transposable elements statistics for the *G. abbreviatus* genome.

Non-coding gene annotation. Non-coding RNAs (ncRNAs) are RNA molecules that do not encode proteins, including transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), microRNAs (miRNAs) and small nuclear RNAs (snRNAs) were annotated using different strategies according to their sequence and structural characteristics. tRNAs were identified with tRNAscan-SE (v1.3.1; parameters: -q)[33]. rRNA genes were detected with barrnap (v0.9; parameters:–quiet–kingdom), taking advantage of the high sequence conservation of rRNAs[34]. Considering the repetitive nature of rRNA gene clusters, the annotation results were further evaluated to reduce potential omissions caused by assembly complexity. miRNAs and snRNAs were predicted with Infernal (parameters: cmscan;–rfam–nohmmonly) against Rfam (v14.8) covariance models[35]. In total, 5,486 tRNAs, 2,417 rRNAs, 541 miRNAs and 1,060 snRNAs were annotated (Table 7).

	Gene set	Gene number	Average Gene length (bp)	Average CDS length (bp)	Average Exons per gene	Average Exon length (bp)	Average Intron length (bp)
De novo	Genscan	26,968	14,909	1,637	8.79	186.22	1,704
	AUGUSTUS	38,827	8,409	1,124	6.17	182.13	1,408
Homolog	*Larimichthys crocea*	46,535	20,471	2,194	12.60	174.20	1,576
	Chanos chanos	32,079	13,829	1,766	10.21	173.00	1,310
	Danio rerio	58,795	17,559	2,084	11.44	182.10	1,482
	Colossoma macropomum	45,065	15,836	1,957	11.00	177.85	1,387
Liftoff	*Chanos chanos*	10,340	7,877	928.03	3.30	280.83	3,015
	Danio rerio	4,961	6,487	835.79	2.12	393.54	5,029
Trans ORF	RNAseq	8,481	11,023	1,489	10.17	298.09	871.18
BUSCO		3,631	9,270	1,645	10.28	160.06	821.71
MAKER		22,133	12,359	1,658	10.10	248.26	1,083
HiFAP		22,396	12,504	1,727	9.97	236.51	1,131

Table 5. Statistics of protein-coding gene models from different evidence sources used for gene annotation in the *G. abbreviatus* genome.

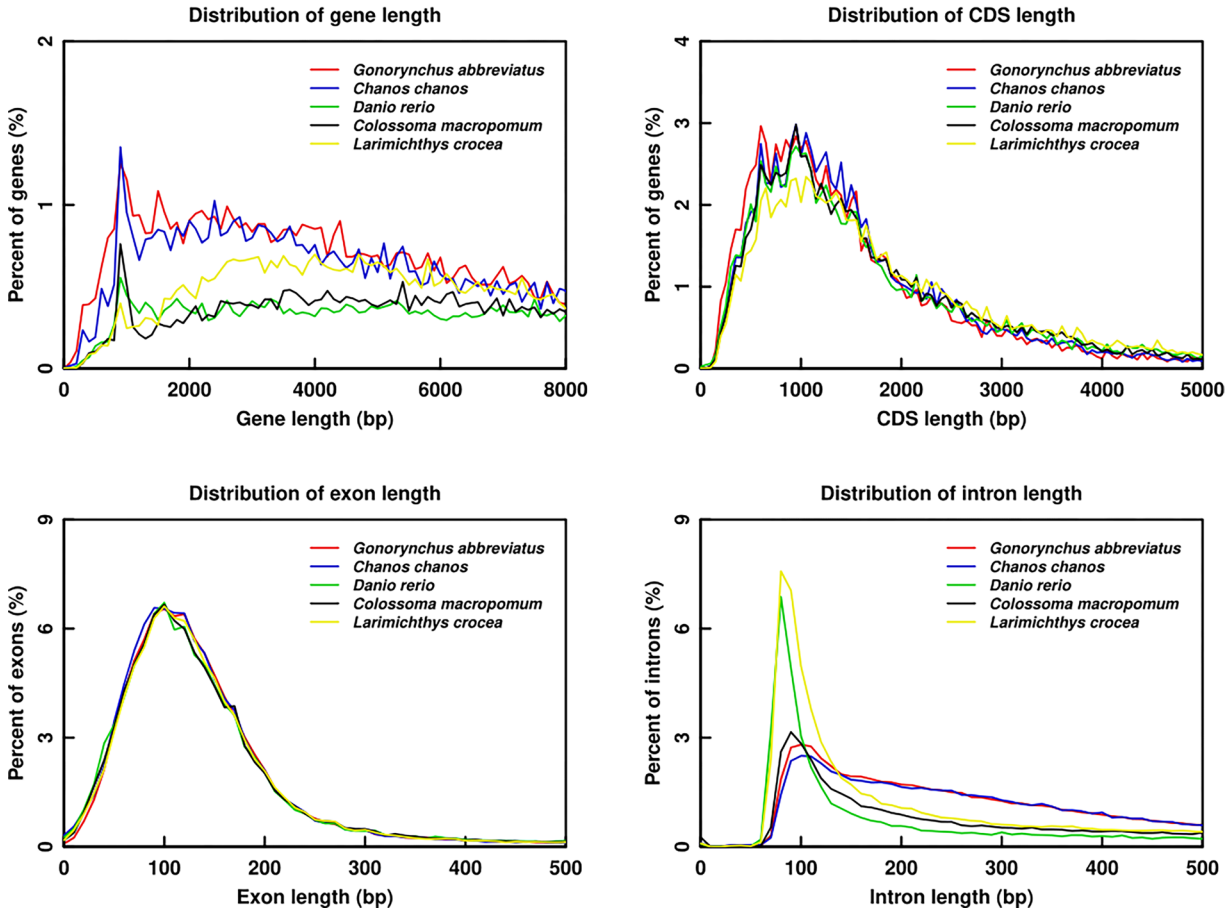

Fig. 4 Comparison of protein-coding genes annotation quality. Five species (*G. abbreviatus*, *Chanos chanos*, *Danio rerio*, *Colossoma macropomum*, and *Larimichthys crocea*) were examined to compare the lengths of the gene, CDS, exon, and intron.

Data Record

Multiple-platform raw sequencing reads have been deposited in the NCBI Sequence Read Archive (SRA), including DNBSEQ reads (SRR37624023), PacBio reads (SRR37624020, SRR37624021, and SRR37624022), Hi-C reads (SRR37624019), and transcriptome reads (SRR37624018)[36–41]. The genome assembly has been deposited in GenBank under accession GCA_054828645.1 (assembly version: ASM5482864v1)[42].

			Gene		mRNA	
			Number	Percent (%)	Number	Percent (%)
Total			22,396	100.00	28,254	100.00
	Annotated		21,429	95.68	27,070	95.81
		NR	21,325	95.22	26,942	95.36
		SwissProt	19,756	88.21	24,947	88.30
		TrEMBL	21,323	95.21	26,939	95.35
		KOG	18,280	81.62	23,175	82.02
		TF	5,016	22.40	6,236	22.07
		InterPro	20,719	92.51	25,858	91.52
		GO	16,026	71.56	19,829	70.18
		KEGG_ALL	21,266	94.95	26,865	95.08
		KEGG_KO	15,391	68.72	19,709	69.76
		Pfam	19,638	87.69	24,316	86.06
	Unannotated		967	4.32	1,184	4.19

Table 6. Putative protein-coding gene functional annotations of the *G. abbreviatus* genome.

Type		Copy number	Average length (bp)	Total length (bp)	Percentage of genome (%)
miRNA		541	75	40,452	0.00664
tRNA		5,486	75	413,034	0.0678
rRNA	Total	2,417	339	818,193	0.134306
	18S	86	1,809	155,561	0.025535
	28S	117	3,425	400,746	0.065782
	5.8S	82	153	12,543	0.002059
	5S	2,132	117	249,343	0.04093
snRNA	Total	1,060	145	153,322	0.025168
	CD-box	178	127	22,627	0.003714
	HACA-box	77	150	11,531	0.001893
	splicing	795	147	116,927	0.019194
	scaRNA	10	224	2,237	0.000367

Table 7. Statistics of the noncoding RNA in the *G. abbreviatus* genome.

Technical Validation

Evaluating quality of the DNA and RNA. Prior to the genome sequencing, we used the NanoDrop 2000 Spectrophotometer (Thermo Fisher Scientific, San Jose, CA, USA) and Qubit 3.0 Fluorometer (Thermo Fisher Scientific, San Jose, CA, USA) to determine the quality (OD260/280 and OD260/230) and concentration of the DNA and RNA samples to ensure the accuracy of sequencing data. We also used the agarose gel electrophoresis and Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, California, USA) to determine the integrity of the DNA and RNA samples.

Quality control of sequencing datasets. Raw DNBSEQ short reads may contain adapter contamination, low-quality bases, and ambiguous nucleotides (Ns), which can negatively affect downstream analyses. Therefore, raw short reads were processed using fastp (v0.23.2) to generate clean reads. Specifically, fastp was used to (i) remove low-quality, too-short, or N-rich reads; (ii) trim low-quality bases from both 5′ and 3′ ends using a sliding-window–based quality evaluation; (iii) automatically detect and remove adapter sequences; (iv) correct bases in overlapped regions of paired-end reads according to base quality; (v) trim artificial 3′ polyG/polyX tails associated with sequencing chemistry; and (vi) perform deduplication to remove PCR duplicates. The resulting clean reads retained the same format as the raw data and were used for subsequent analyses[43].

PacBio HiFi reads were generated with high per-base accuracy and were used directly for genome assembly. Read length distribution, total sequencing yield, and sequencing depth were evaluated to ensure sufficient data quality and coverage for *de novo* assembly, minimising the need for extensive read-level filtering. Hi-C reads were subjected to basic quality control prior to scaffolding. Clean Hi-C reads were aligned to the contig-level assembly, and mapping statistics were examined, including the proportion of uniquely mapped read pairs and valid Hi-C contacts, to assess the quality of the Hi-C library and its suitability for chromosome-scale scaffolding.

Evaluating quality of the genome assembly and annotation. To evaluate the sequence consistency and assembly quality, the BWA (v0.7.17-r1188) and Minimap2 (v2.24-r1122) with the default parameter, were used to map the short reads from DNBSEQ and HiFi reads from PacBio to the assembled genome,

respectively[44,45]. After these processes, 97.22% of the short reads from DNBSEQ and 97.74% of the HiFi reads from PacBio were aligned, covering 99.90% and 100.00% of the assembled genome, respectively (Table S4).

Moreover, BUSCO (v5.7.1) was employed to evaluate the quality of assembly and annotation using the actinopterygii_odb12 database, which contains 7,207 conserved orthologs[46]. For assembly evaluation, BUSCO was run in genome mode (–miniprot -m geno -f) using the assembled genome as input. A total of 99.07% of BUSCO genes were identified as complete, including 98.32% complete single-copy and 0.75% complete duplicated BUSCOs. In addition, 0.08% and 0.85% of BUSCOs were classified as fragmented and missing, respectively (Table S5). For annotation evaluation, BUSCO was run in protein mode (–miniprot -m prot -f) using the predicted protein sequences as input. 92.66% of BUSCO genes were identified as complete, including 91.69% complete single-copy and 0.97% complete duplicated BUSCOs, while 2.46% were fragmented and 4.88% were missing (Table S6).

Data availability

All data have been deposited at the National Center for Biotechnology Information (NCBI) under BioProject accession PRJNA1398026 and BioSample accession SAMN54439668[27,47,48]. The raw sequencing reads are available in the NCBI Sequence Read Archive (SRA) under Study accession SRP683575, and the genome assembly is available in GenBank under accession GCA_054828645.1[42,49]. Further details are provided in the Data Record section below.

Code availability

No custom code was generated for this study. All analyses were performed using publicly available bioinformatics software according to the developers' manuals and protocols, with parameter settings reported in the Methods section. An overview of the analytical workflow is provided in Figure S1 (Supplementary Materials).

Received: 7 October 2025; Accepted: 19 March 2026;
Published online: 10 April 2026

References

1. Gonorynchus abbreviatus. *Ncbi* https://www.ncbi.nlm.nih.gov/datasets/taxonomy/189504/names/.
2. Grande, T. Revision of the genus gonorynchus scopoli, 1777 (teleostei: ostariophysi). *Copeia* **1999**, 453–469 (1999).
3. Poyato-Ariza, F. J., Grande, T. & Diogo, R. Gonorynchiform interrelationships: historic overview, analysis, and revised systematics of the group. in *Gonorynchiformes and Ostariophysan Relationships* (CRC Press, 2010).
4. Grande, T. & Arratia, G. Morphological analysis of the gonorynchiform postcranial skeleton. in *Gonorynchiformes and Ostariophysan Relationships* (CRC Press, 2010).
5. Miya, M. & Nishida, M. Molecular phylogenetic perspective on the evolution of the deep-sea fish genusCyclothone (stomiiformes: gonostomatidae). *Ichthyol. Res.* **43**, 375–398 (1996).
6. Volff, J.-N. Genome evolution and biodiversity in teleost fish. *Heredity* **94**, 280–294 (2005).
7. Ravi, V. & Venkatesh, B. The divergent genomes of teleosts. *Annu. Rev. Anim. Biosci.* **6**, 47–68 (2018).
8. Glasauer, S. M. K. & Neuhauss, S. C. F. Whole-genome duplication in teleost fishes and its evolutionary consequences. *Mol. Genet. Genomics* **289**, 1045–1060 (2014).
9. Ahmad, S. F., Jehangir, M., Srikulnath, K. & Martins, C. Fish genomics and its impact on fundamental and applied research of vertebrate biology. *Rev. Fish Biol. Fish.* **32**, 357–385 (2022).
10. Phylogenetic relationships and timing of diversification in gonorynchiform fishes inferred using nuclear gene DNA sequences (teleostei: ostariophysi). *Mol. Phylogenet. Evol.* **80**, 297–307 (2014).
11. Hi–C: a comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268–276 (2012).
12. Liu, B. *et al*. Estimation of genomic characteristics by analyzing k-mer frequency in *de novo* genome projects. Preprint at https://doi.org/10.48550/arXiv.1308.2012 (2020).
13. 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).
14. Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinf.* **19**, 460 (2018).
15. Jurka, J. *et al*. Repbase Update, a database of eukaryotic repetitive elements. *Cytogene. Genome Res.* **110**, 462–467 (2005).
16. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
17. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
18. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA1. *J. Mol. Biol.* **268**, 78–94 (1997).
19. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinf.* **6**, 31 (2005).
20. Shumate, A. & Salzberg, S. L. Liftoff: accurate mapping of gene annotations. *Bioinformatics* **37**, 1639–1643 (2021).
21. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
22. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
23. TransDecoder/TransDecoder. TransDecoder (2025).
24. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinf.* **12**, 491 (2011).
25. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).
26. Coudert, E. *et al*. Annotation of biologically relevant ligands in UniProtKB using ChEBI. *Bioinformatics* **39**, btac793 (2023).
27. Wheeler, D. L. *et al*. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **36**, D13–D21 (2008).
28. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
29. Index of /pub/COG/KOG. https://ftp.ncbi.nih.gov/pub/COG/KOG/.
30. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
31. Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. & Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **41**, e121 (2013).

32. Bu, D. *et al*. KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis. *Nucleic Acids Res*. **49**, W317–W325 (2021).
33. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*. **25**, 955–964 (1997).
34. Seemann, T. tseemann/barrnap. (2026).
35. Griffiths-Jones, S. *et al*. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res*. **33**, D121–D124 (2005).
36. *NCBI Sequence Read Archive* https://www.ncbi.nlm.nih.gov/sra/SRR37624018 (2026).
37. *NCBI Sequence Read Archive* https://www.ncbi.nlm.nih.gov/sra/SRR37624019 (2026).
38. *NCBI Sequence Read Archive* https://www.ncbi.nlm.nih.gov/sra/SRR37624020 (2026).
39. *NCBI Sequence Read Archive* https://www.ncbi.nlm.nih.gov/sra/SRR37624021 (2026).
40. *NCBI Sequence Read Archive* https://www.ncbi.nlm.nih.gov/sra/SRR37624022 (2026).
41. *NCBI Sequence Read Archive* https://www.ncbi.nlm.nih.gov/sra/SRR37624023 (2026).
42. *NCBI Assembly* https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_054828645.1 (2026).
43. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
44. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
45. Li, H. & Durbin, R. Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
46. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO update: novel and streained workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol*. **38**, 4647–4654 (2021).
47. *NCBI BioProject* https://www.ncbi.nlm.nih.gov/bioproject/1398026 (2025).
48. *NCBI BioSample* https://www.ncbi.nlm.nih.gov/biosample/SAMN54439668 (2025).
49. *NCBI Sequence Read Archive* https://www.ncbi.nlm.nih.gov/sra/SRP683575 (2026).

Acknowledgements

We thank the reviewers and the handling editor for their valuable comments and suggestions, which greatly improved the quality of this manuscript. This research was financially supported by National Key R & D Program of China (2022YFF0802204).

Author contributions

Linlin Zhao: Writing – original draft, Methodology, Software, Validation, Data curation, Funding acquisition; Xiaotong Zhu: Methodology, Writing – review & editing; Min Zhou: Methodology, Writing – review & editing; Ruizhi Wang: Methodology, Writing – review & editing; Na Song: Methodology, Writing – review & editing; Bailin Cong: Methodology, Writing – review & editing; Wenbo Zhu: Methodology, Project administration, Conceptualisation, Writing – review & editing.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-026-07121-6.

Correspondence and requests for materials should be addressed to B.C. or W.Z.