scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A telomere-to-telomere reference genome assembly of the *Hypomesus nipponensis*

Yanfeng Zhou[1] ✉, Di'an Fang[1], Yang You[1], Fujiang Tang[2], Yulin Bai[1], Minying Zhang[1], Xuemei Li[3], Guoping Deng[4] & Dongpo Xu[1] ✉

A small cold-water teleost endemic to Northeast Asia, *Hypomesus nipponensis* possesses a short lifecycle, high fecundity, and rapid population growth, with extensive introductions for aquacultural purposes across East Asia. In this study, we generated a gap-free, telomere-to-telomere (T2T) genome assembly of *H. nipponensis* using a combined sequencing strategy, incorporating MGI short reads, PacBio High-Fidelity (HiFi) reads, Oxford Nanopore Technologies (ONT) ultra-long reads, and Hi-C data. The final assembly spans 526.31 Mb with a contig N50 of 20.23 Mb, and all genomic sequences were successfully anchored to 28 pseudochromosomes. BUSCO assessment (Actinopterygii_odb10) confirms 98.19% completeness, including 3,548 single-copy and 26 duplicated orthologs out of 3,640 conserved genes. Repeat elements account for 39.17% (206.18 Mb) of the genome, and 31,310 protein-coding genes are annotated. This gap-free T2T assembly resolves previously uncharacterized genomic regions, providing a high-quality reference for molecular breeding, evolutionary analyses of the *Hypomesus* genus, and functional investigations into adaptive traits of cold-water fishes.

Background & Summary

Hypomesus nipponensis (NCBI Taxonomy ID: 182223), an anadromous small cold-water fish classified under the genus *Hypomesus* (family Osmeridae), exhibits a short life cycle, high fecundity, and rapid population growth—adaptive traits that support its colonization of diverse water bodies and survival across heterogeneous habitats[1]. Prior to the 1980s, Hypomesus nipponensis (Japanese smelt) was introduced to northeastern China, with Shuifeng Reservoir—the largest reservoir in Northeast China—recognized as one of its core introduction sites[2]. As a highly dispersive species, this fish has now established a distribution range covering the entire Northeast Asia, spanning China, Japan, and the Korean Peninsula[3].

Whole genome information serves as the foundation for investigating biological characteristics. To date, several genomic resources for *H. nipponensis* have been reported. In 2019, the complete mitochondrial genome of *H. nipponensis* was decoded[4]. Subsequently, a draft genome was generated in 2021, with a contig N50 of 464,523 bp[5]. Most recently, a chromosome-level genome assembly (designated HNIP-V2) with a contig N50 of 8.19 Mb was published[6]. These studies have provided critical genetic resources and established a robust foundation for breeding programs and biological research on *H. nipponensis*. However, these present genome assemblies have been limited by numerous gaps, particularly in repetitive sequence-rich regions such as telomeres and centromeres. Telomeric and centromeric DNA sequences are predominantly composed of satellite DNA and are known to evolve rapidly in eukaryotic genomes[7,8]. With advancements in genome sequencing technologies and assembly methodologies, gap-free telomere-to-telomere (T2T) genome assemblies have now become achievable, enabling the characterization of nearly the entire genome. Pacific BioSciences (PacBio) HiFi reads can resolve complex genomic regions, while ONT ultra-long reads facilitate the resolution of tandem duplications[9,10]. Hifiasm, a high-performance assembly tool, has been successfully applied to gap-free T2T genome assembly in various fish species, including the Yangtze finless porpoise (*Neophocaena asiaeorientalis*)[11], *Neosalanx taihuensis*[12],

[1]Key Laboratory of Freshwater Fisheries and Germplasm Resources Utilization, Ministry of Agriculture and Rural Affairs, Freshwater Fisheries Research Center, Chinese Academy of Fishery Sciences, Wuxi, 214081, China. [2]Heilongjiang River Fisheries Research Institute, Chinese Academy of Fishery Sciences, Harbin, 150070, China. [3]Yangtze River Fisheries Research Institute, Chinese Academy of Fishery Sciences, Wuhan, China. [4]Dalian Ocean University, Dalian, 116023, China. ✉e-mail: zhouyf@ffrc.cn; xudp@ffrc.cn

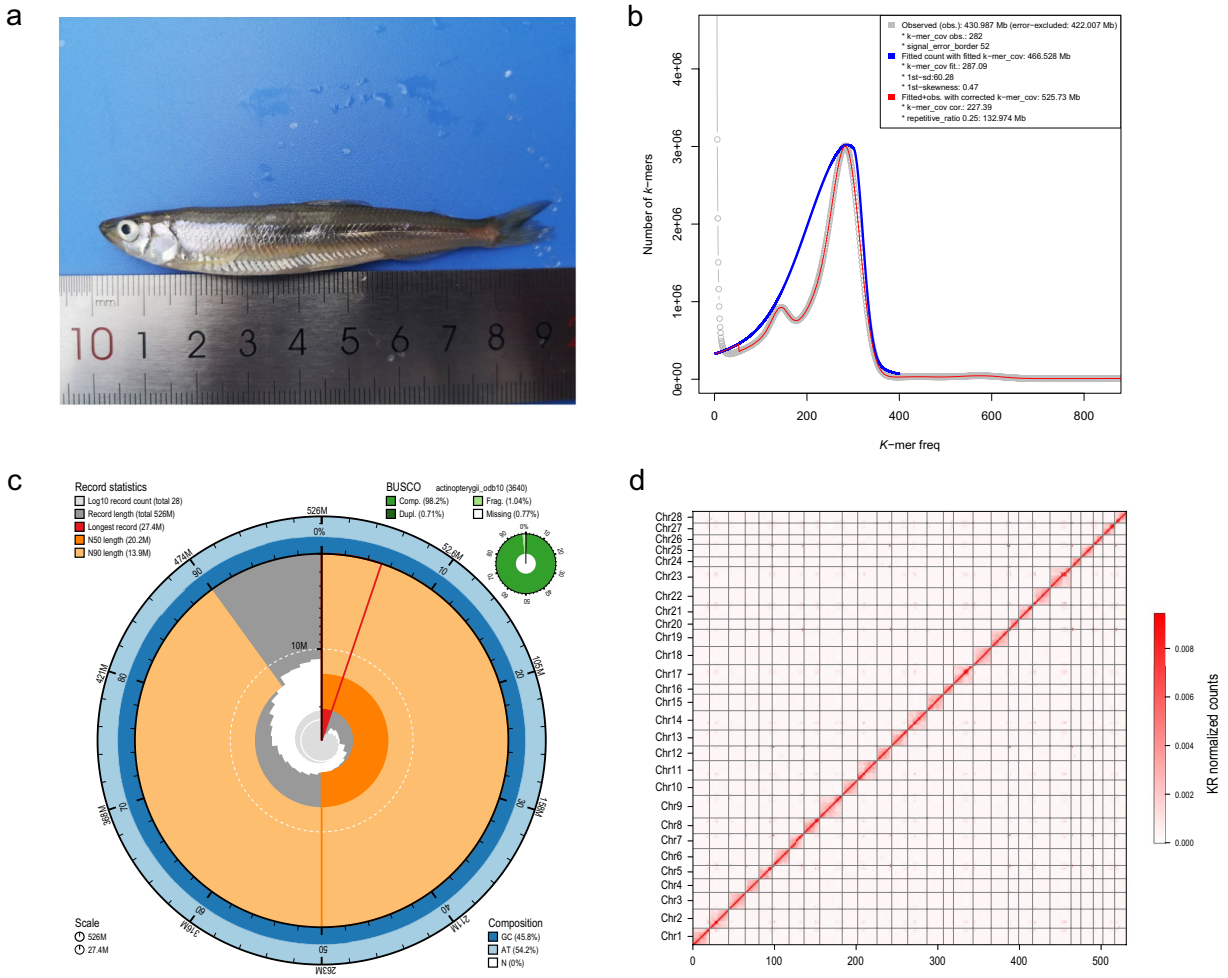

Fig. 1 A T2T genome assembly of *H. nipponensis*. (**a**) An image of the sequenced fish. (**b**) *K*-mer frequency distribution estimated. The observed *K*-mer (raw *K*-mer) frequencies (in grey), fitted *K*-mer frequencies (in blue) with skew normal distribution model, and overall fitting (in red) that concatenated observed and fitted *K*-mer frequencies. Genome size estimate for 19-mer: 525,729,808 bp. (**c**) Snail plot showing the features of the assembled *H. nipponensis* genome. The contiguity and completeness of the *H. nipponensis* genome assembly after contamination screening is plotted as a circle that represents the full length of the assembly (~526.31 Mb). The N50 (20.23 Mb) is highlighted in dark orange and the N90 (13.86 Mb) in light orange. The longest contig was 27.41 Mb (highlighted in red). The assembly has a uniform GC content of 45.85% and the BUSCO scores are shown in the top right corner in green. (**d**) Hi-C assembly of chromosome interactive heat map. The abscissa and ordinate represent the order of each bin on the corresponding chromosome group. The colour block illuminates the intensity of interaction from white (low) to red (high).

Asian icefish (*Protosalanx chinensis*)[13], and *Siniperca roulei*[14]. Notably, its algorithm has recently been updated to specifically support T2T assembly using ONT data alone[15].

In this study, we report the first gap-free T2T reference genome for *H. nipponensis* (designated HNIP-T2T), generated using multiple assembly strategies and integrating HiFi reads, ONT reads, MGI short reads, and chromatin conformation capture (Hi-C) data. The HNIP-T2T assembly spans approximately 526.31 Mb with an N50 of 20.23 Mb. Gene annotation identified 31,310 protein-coding genes, 97.67% of which were annotated in public biological databases. This high-quality, gap-free genome assembly will serve as an important resource for investigating the reproductive biology and ecological adaptability of *H. nipponensis*.

Methods

Sample collection and sequencing. In September 2024, a healthy *H. nipponensis* was collected from Shuifeng Reservoir on the Yalu River in Liaoning Province, China (Fig. 1a). High-quality, high-molecular-weight genomic DNA (gDNA) was extracted from muscle tissue using the cetyltrimethylammonium bromide method[16]. DNA purity was assessed through 1% agarose gel electrophoresis and quantified using a NanoDrop™ One UV-Vis spectrophotometer (Thermo Fisher Scientific, USA). DNA concentration was further determined using a Qubit 4.0 fluorometer (Invitrogen, USA). Following quality assessment, paired-end sequencing was performed

Types	Reads Number	Total Length (Gb)	Genome Depth*	N50 Length of Reads (bp)
MGI raw reads	959,642,776	143.95	273.50	150
Hi-C raw data	1,377,792,170	206.67	392.67	150
PacBio HiFi reads	2,150,973	36.32	69.01	17,253
ONT reads	632,072	28.63	54.39	83,817

Table 1. Summary of DNA sequencing data of *H. nipponensis* genome. Note: *Estimated based on the assembly size of 526.31 Mb.

Method	HiFiasm			NextDenovo
Read type	PacBio	PacBio + error-corrected ONT	ONT	ONT
N50 (bp)	666,015	4,460,278	18,759,114	15,907,864
Largest contig	10,531,176	20,739,669	27,407,527	24,001,982
Total size (bp)	505,311,569	554,942,883	774,050,513	516,291,601
Total number	2,012	566	76	72
GC rate (%)	45.6	45.7	45.9	45.8

Table 2. Assembly statistics using two different assembly software.

on the DNBSEQ-T7 platform (MGI, Shenzhen, China), generating 143.95 Gb of raw reads (Table 1). Quality control of the sequencing data was conducted using fastp (v0.23.2)[17] with default settings.

For long-read sequencing, a SMRTbell library was constructed and sequenced using the PacBio Revio system (Pacific Biosciences, USA). Following preprocessing with the CCS program[18], 36.32 Gb of high-quality Circular Consensus Sequencing (CCS) reads were generated, corresponding to a sequencing depth of ~69.01 × with an N50 value of 17,253 bp (Table 1).

ONT technology was applied by constructing an ultra-long library and then sequencing of one flow cell on a PromethION platform (Oxford Nanopore Technologies Co., UK). The raw reads were first filtered to remove bases with quality value (QV) below 7. Adapter sequences were then trimmed using Porechop (https://github.com/rrwick/Porechop). Finally, reads in which fewer than 90% of bases achieved QV ≥ 7 were removed using Filtlong (https://github.com/rrwick/Filtlong). Finally, we obtained a total of 28.63 Gb clean reads, with an N50 length of 83.82 kb.

For Hi-C sequencing, we extracted gDNA, digested chromatin using the restriction enzyme MboI, and then conducted proximity ligation according to protocols outlined in previous studies[19]. In brief, gDNA was cross-linked, digested, biotin-labeled, ligated, and fragmented to 350 bp, followed by purification with streptavidin magnetic beads. Library quality and insert size were assessed by using a Qubit 3.0 Fluorometer and an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA), respectively. Libraries were sequenced on DNBSEQ-T7 platform (MGI, Shenzhen, China), generating ~206.67 Gb of 150 bp paired-end reads (Table 1).

Genome size estimation. Through *K*-mer analysis (K = 19) of MGI short clean reads with Jellyfish (v2.3.0)[20], an overall *H. nipponensis* genome size of 525.73 Mb was estimated using findGSE (v1.94)[21] (Fig. 1b).

***De novo* genome assembly.** Initially, contig assembly was performed using the HiFiasm (v0.25.0-r726)[22] software based on three distinct datasets: 1 > PacBio HIFI data; 2 > a combined dataset of corrected ONT data, PacBio HIFI data and Hi-C data; 3 > raw ONT data (without error correction). Assembly using PacBio HiFi reads alone yielded 2,012 contigs, with a total length of 505.31 Mb and an N50 value of 666,015 bp. When combining error-corrected ONT reads with PacBio HiFi reads, 566 contigs were generated, with a total length of 554.94 Mb and an N50 of 4.46 Mb (Table 2). Assembly using raw ONT reads alone resulted in 76 contigs, with a total length of 774.05 Mb and an N50 of 18.76 Mb (Table 2). Additionally, ONT reads were error corrected using NextDenovo (v2.5.0)[23], producing 6.23 Gb of error-corrected reads with an N50 length of 143,361 bp. These error-corrected ONT reads were subsequently assembled independently using NextDenovo (v2.5.0)[23], yielding an assembly with a total length of 516.29 Mb and an N50 of 15.91 Mb (Table 2). Based on the core consideration of sequence contiguity, the assembly result generated by HiFiasm using raw ONT data (without error correction) was ultimately selected as the genome scaffold for subsequent analyses.

The contigs generated from ONT-only assembly using HiFiasm (v0.25.0-r726)[22] were polished by pilon (v1.24)[24] using clean short reads from MGI sequencing. Purge-Haplotigs (v1.1.2)[25] was employed to reduce haplotypic duplication, thereby refining the assembly continuity and haploidy. Following the generation of non-redundant contigs, Hi-C clean reads were mapped to this assembly using Bowtie2 (v2.2.5)[26] with parameters:–very-sensitive -L 30–score-min L,-0.6,-0.2–end-to-end –reorder, and then effective linkage products were detected using HiC-Pro (v2.8.1)[27] under default settings, retaining only valid contact pairs to support the anchoring of contigs to chromosomes. To orient, order, and cluster contigs into pseudochromosomes, we applied Juicer (v1.5)[28] and 3D-DNA (v170123)[29]. Visualization and manual corrections were performed with Juicebox (v1.11.08)[30] to adjust mis-assemblies and eliminate redundant contigs. Ultimately, 28 pseudo-chromosomes were obtained with only six gaps remaining (Table 3). The longest and shortest pseudo-chromosomes measured

Pseudomolecule	Length (bp)	GC content (%)	Gap number
Chr1	20,312,735	45.05	0
Chr2	23,318,322	45.70	2
Chr3	20,232,964	46.27	0
Chr4	16,615,911	46.50	0
Chr5	16,143,452	46.16	0
Chr6	20,584,329	45.48	0
Chr7	18,284,807	46.52	0
Chr8	18,984,847	45.38	0
Chr9	27,401,941	45.00	0
Chr10	18,755,640	45.73	0
Chr11	23,418,985	45.21	0
Chr12	17,823,600	46.39	2
Chr13	19,361,263	45.62	0
Chr14	23,418,982	45.39	0
Chr15	19,835,662	46.13	1
Chr16	12,282,191	46.67	0
Chr17	23,852,489	45.79	0
Chr18	22,072,218	45.30	0
Chr19	20,746,590	45.66	0
Chr20	12,395,499	47.30	0
Chr21	16,873,045	46.14	0
Chr22	21,168,038	45.81	1
Chr23	25,749,877	45.10	0
Chr24	11,805,118	47.84	0
Chr25	15,192,889	46.86	0
Chr26	11,481,171	46.45	0
Chr27	14,354,222	44.66	0
Chr28	13,855,024	46.88	0

Table 3. Pseudo-chromosome length statistics after Hi-C assisted assembly.

Item	HNIP-T2T (This study)	HNIP-V2[1]
Size of assembly (Mb)	526.31	532.61
Contig N50 (Mb)	20.23	8.19
Gap number	0	189
GC content (%)	45.85	45.84
Genome complete BUSCOs (%)	98.19	96.7
Quality value	35.11	39.41
MGI clean reads mapping rate (%)	99.72	—
ONT reads mapping rate (%)	99.86	—
HiFi reads mapping rate (%)	99.97	—
Repetitive sequences (%)	39.17	33.59
Number of protein-coding genes	31,310	27,876

Table 4. Summary statistics of *H. nipponensis* assembly. Note: The lineage dataset used in BUSCO is actinopterygii_odb10.

27.40 Mb and 11.48 Mb, respectively. This chromosome number aligns with the count reported in the previously published HNIP-V2 assembly[6] and is consistent with the karyotype of *Hypomesus olidus* (2n = 56)[31].

To achieve a gap-free and telomere-to-telomere (T2T)-level assembly, LR_GapCloser (v1.0)[32] was sequentially employed to fill gaps using PacBio HiFi reads and error-corrected ONT long reads, with the following parameters: -m 1000000 -v 10000 -r 3. The resulting HNIP-T2T assembly comprises 28 anchored pseudochromosomes, with a total length of 526.31 Mb (Table 4). The N50 value of these anchored chromosomes was increased to 20.23 Mb (Table 4 and Fig. 1c). Notably, the Hi-C interaction heatmap exhibited high consistency across all pseudochromosomes, confirming the accuracy of sequencing data, contig ordering, and orientation in the HNIP-T2T assembly (Fig. 1d). The chromosome order and orientation of the HNIP-T2T assembly

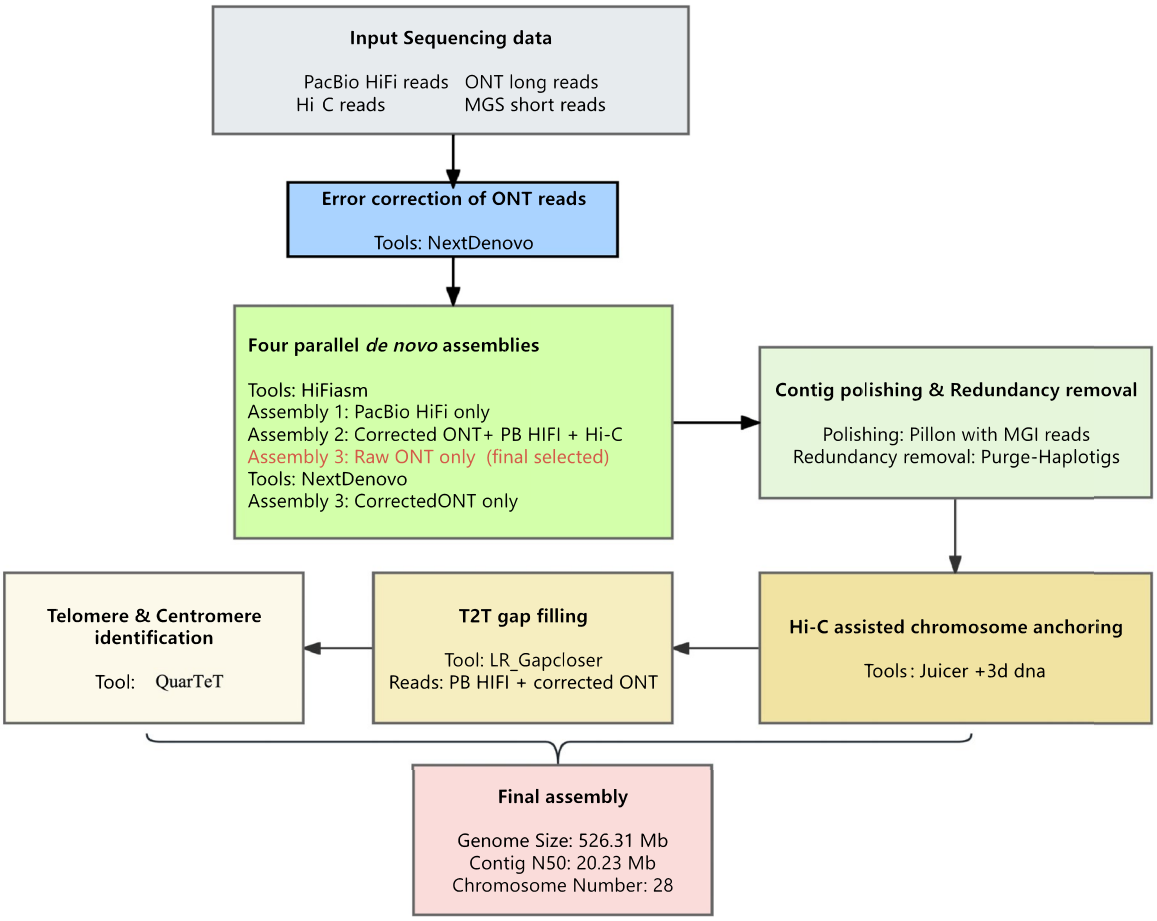

Fig. 2 Overview of the *de novo* genome assembly pipeline for *H. nipponensis* (T2T-level).

were adjusted with reference to the reference genome of *Danio rerio* (zebrafish; GenBank assembly accession: GCF_000002035.6), ensuring comparability with the genomic structure of this model species.

The detailed assembly pipeline is illustrated in Fig. 2.

Identification of centromere and telomere sequences. Using the QuarTeT software[33], we identified centromere and telomere sequences in the HNIP-T2T genome. QuarTeT's centromere prediction relies on three integrated signals: (1) tandem repeat enrichment (Tandem Repeats Finder parameters: match = 2, mismatch = 7, indel = 7, minimum score = 50); (2) CENH3 homolog co-localization; (3) low recombination/high divergence signatures from read depth analysis—this strategy compensates for the lack of *H. nipponensis* karyotypic data. All 28 pseudochromosomes harbored intact telomeres and centromeres, including 56 telomeres and 28 centromeres (average length: 316,527 bp; Fig. 3). Centromere lengths varied significantly, ranging from 104,388 bp (pseudochromosome 15) to 1,690,782 bp (pseudochromosome 7). Future validation via fluorescence *in situ* hybridization (FISH, for chromosomal localization) and CENH3-targeted chromatin immunoprecipitation sequencing (ChIP-seq, for functional verification) will confirm centromere positions, addressing the current gap in *H. nipponensis* karyotypic research.

Repeat element annotation. In HNIP-T2T, repetitive elements were identified through integration of *de novo* and homology-based annotation methods. The homology-based blast was performed against the RepBase database (http://www.girinst.org/repbase/)[34] using RepeatMasker (v4.0.7)[35] and Proteinmask software for known repeat elements. For *de novo* annotation, we firstly used LTR_FINDER (v1.06)[36] and RepeatModeler (v1.0.4)[37] to construct a *de novo* repeat library. This library was then used to predict repetitive elements with RepeatMasker (v4.0.7)[35] under default parameters. Additionally, Tandem Repeat Finder (v4.10.0)[38] was applied to identify tandem repeats using settings: 2 7 7 80 10 50 2000 -d -h. In detail, a total of 206.18 Mb (39.17%) of repetitive sequences were obtained. The proportion of repetitive sequences is higher than that in HNIP-V2 (33.59%)[6]. Among the interspersed repeats, DNA transposons were the most abundant type, representing 16.95% of the genome (Table 5).

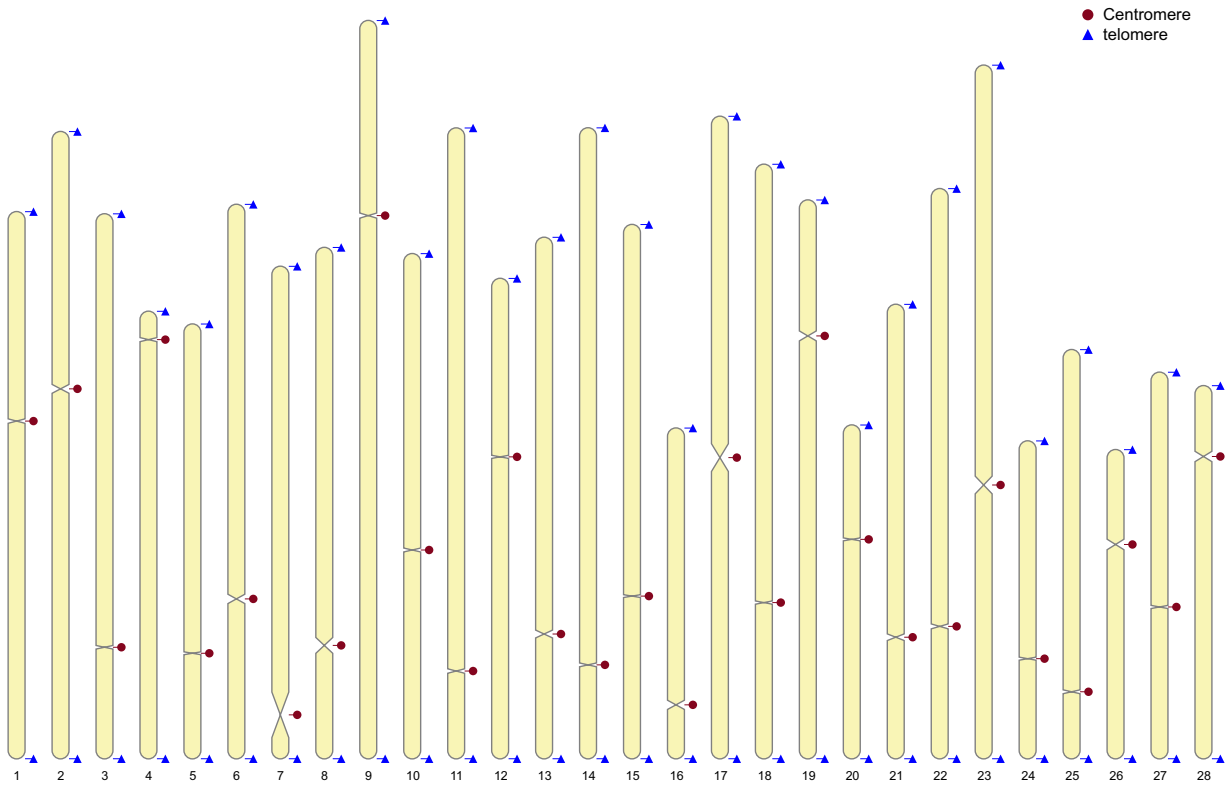

Fig. 3 Telomere and centromere locations in the *H. nipponensis* genome. The triangle represents the telomere region, and the circle represents the centromere region.

Type	Repbase TEs	TE protiens	*De novo*	Combined TEs
	% in genome	% in genome	% in genome	% in genome
DNA	9.22	0.12	12.12	16.95
LINE	3.14	1.44	10.83	12.21
SINE	1.31	0.00	4.38	4.53
LTR	2.28	0.53	8.32	9.70
Other	0.00	0.00	0.00	0.00
Unknown	0.00	0.00	0.87	0.87
Total	14.06	2.09	29.93	32.20

Table 5. Statistics of interspersed repetitive sequences in *H. nipponensis* assembly. Note: This statistical table does not contain Tandem Repeats, some elements may partly include another element domain. Combined: the non-redundant consensus of all repeat prediction/classification methods employed. LINE, long interspersed nuclear elements; SINE, short interspersed nuclear elements; LTR, long terminal repeat.

Gene prediction and functional annotation. Gene structure annotation was performed following the established methodology from pig pan-genome research[39]. For transcriptome-based annotation, approximately 33.52 Gb of RNA-seq data from muscle tissues were mapped to the HNIP-T2T assembly using HISAT2 (v2.2.1)[40] with the following parameters:–sensitive–no-discordant–no-mixed -I 1 -X 1000–max-intronlen 1000000. The unique genome mapping rate ranged from 90.52% to 91.17% (Table 6). Subsequently, transcript assembly was performed using Stringtie (v1.2.2)[41] (parameters: -f 0.3 -j 3 -c 5 -g 100 -s 10000). Coding sequences (CDSs) were identified using TransDecoder (v5.7.1). Genes with complete structures were selected, with only the longest transcript retained for each gene. Single-exon genes were included only if a structural protein domain was detected. We excluded genes with ≥80% overlap between gene regions and repeat sequences, yielding a final transcriptome-derived candidate gene set. For the homology prediction, genome sequences and annotation files were retrieved from five representative species: *Danio rerio* (zebrafish; GCF_000002035.6), HNIP-V2[6], *Hypomesus transpacificus* (GCF_021917145.1), *Neosalanx taihuensis*[12], and *Protosalanx chinensis*[13]. Leveraging these RNA-seq and homology data, CDSs were predicted with GeMoMa (v1.9)[42]. Genes derived from transcriptome data but absent from homology predictions were incorporated into the gene set. Finally, untranslated

Sample	Raw reads	Raw bases (bp)	Clean reads	Clean bases (bp)	Clean reads Q20 (%)	HNIP-T2T unique mapping rate (%)	HNIP-V2 unique mapping rate (%)
SF01C	43,780,490	6,567,073,500	42,525,218	6,378,782,700	97.49	90.66	90.11
SF01D	44,380,700	6,657,105,000	42,996,368	6,449,455,200	97.42	90.52	89.95
SF01E	43,542,798	6,531,419,700	42,102,588	6,315,388,200	97.80	90.67	90.15
SF01F	46,188,238	6,928,235,700	44,386,920	6,658,038,000	97.88	91.17	90.57
SF01G	52,906,478	7,935,971,700	51,481,516	7,722,227,400	97.83	90.94	90.36

Table 6. Summary of RNAseq sequencing data of *H. nipponensis* genome.

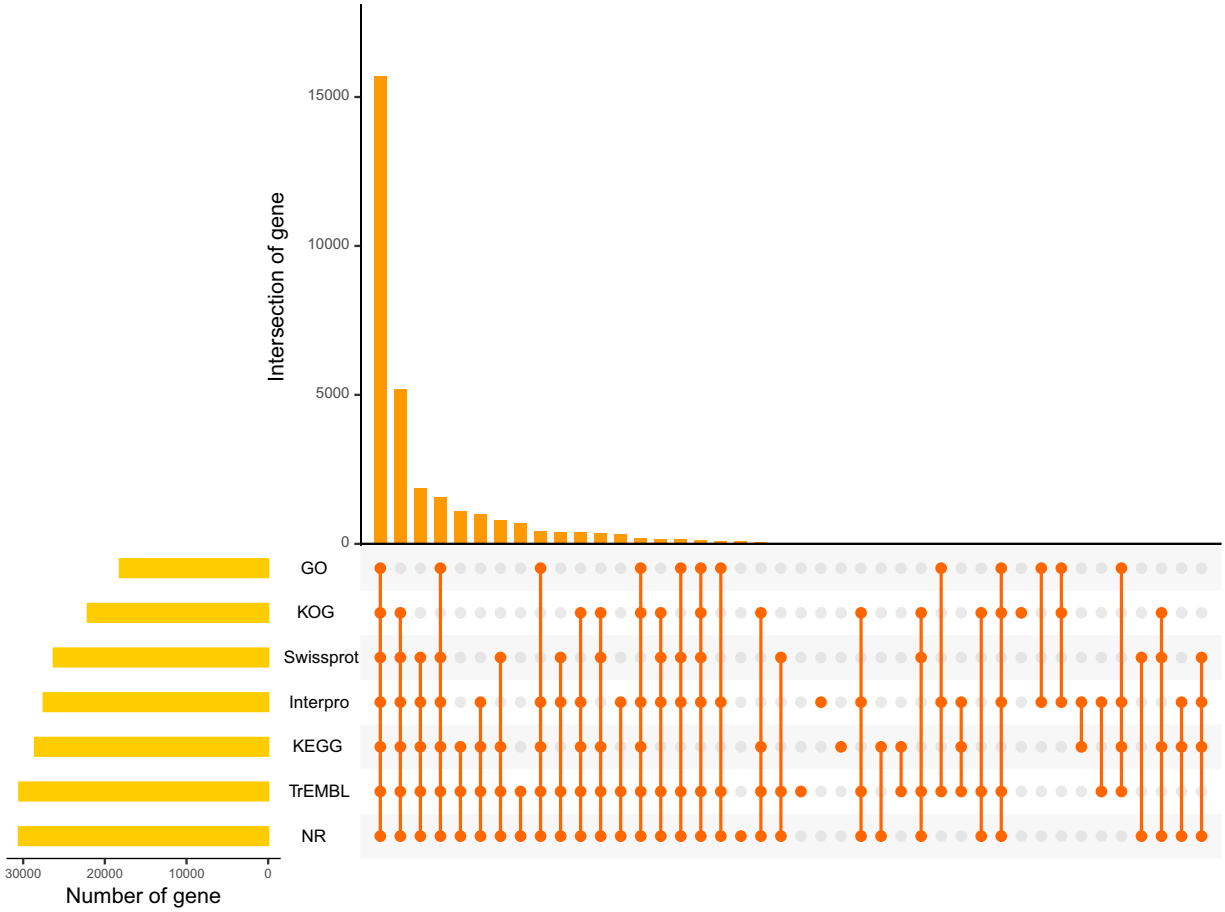

Fig. 4 UpSetR plot showing distribution of gene function annotation. Note: NR, Non-Redundant Protein Sequence Database; Swissport, Swiss-Prot Protein Knowledgebase; KEGG, Kyoto Encyclopedia of Genes and Genomes; KOG, Eukaryotic Orthologous Groups; TrEMBL, Translation of European Molecular Biology Laboratory; Interpro, Integrative Protein Signature Database; GO, Gene Ontology.

regions and alternative splicing variants were annotated using the Program to Assemble Spliced Alignment (v2.4.1)[43]. The final comprehensive gene set comprised 31,310 genes, with a mean of 8.31 exons per gene, an exon length of 191.90 bp, and a CDS length of 1,593.89 bp.

The protein-coding genes were functionally annotated by aligning them with several routine protein databases. Briefly, amino-acid sequences were aligned to SwissProt[44], Kyoto Encyclopedia of Genes and Genomes (KEGG)[45], Eukaryotic Orthologous Groups (KOG)[46], and the NCBI nonredundant database (NR) using the Diamond (v2.1.10)[47] with an E-value cutoff of 1e-05. Protein domains were identified using the InterProScan (v5.30)[48] program, and Gene Ontology (GO) terms for each gene were also extracted through InterProScan. Overall, 30,582 genes (97.67%) were functionally annotated (Fig. 4).

Ethics declarations. Both the sampling procedure and experimental workflow were conducted in strict accordance with the guidelines of the Animal Ethics Committee of the Institute of Hydrobiology, Chinese Academy of Sciences, and have obtained its official approval (Approval Number: IHB1LL12024044).

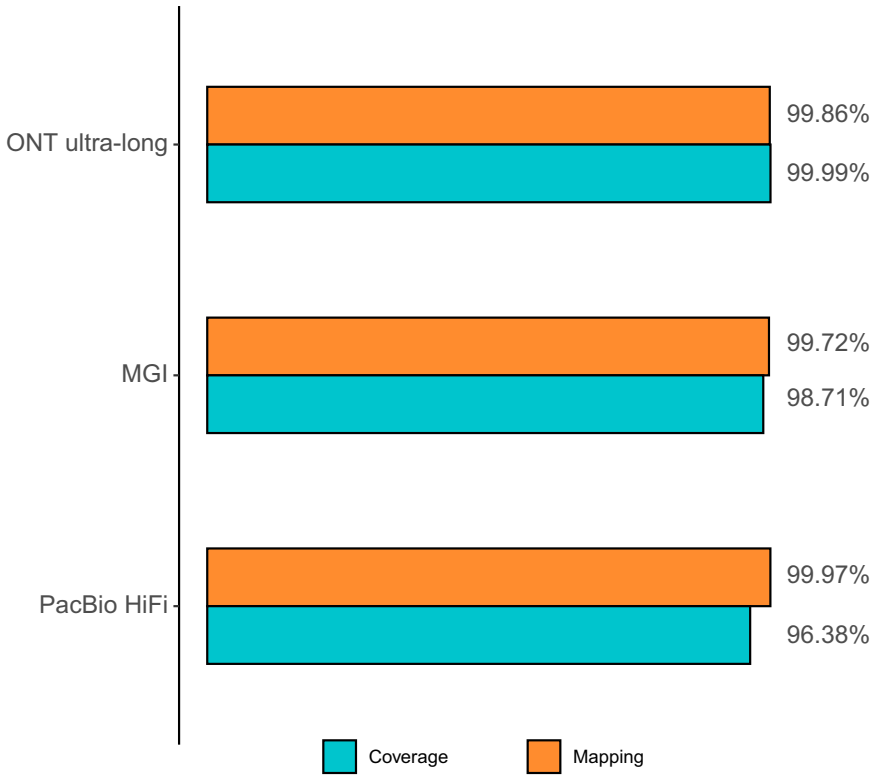

Fig. 5 Mapping rate and coverage of reads from different sequencing platforms.

Data Records

The sequencing data of *Hypomesus nipponensis* presented in this study have been deposited to the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) database under accession number PRJNA1282796[49]. This includes short-read data [RNA-seq data: SRR34259912–SRR34259916; DNA survey data: SRR34259908; and Hi-C data: SRR34259911] and long-read data [Oxford Nanopore Technology (ONT) data: SRR34259910; and PacBio HiFi data: SRR34259909]. The final genome assembly is available under the accession numbers JBTLQK000000000 and GCA_054491055.1[50]. Furthermore, the final genome assembly, annotated coding sequences, and protein sequences are available at Figshare[51].

Technical Validation

To assess the accuracy and quality of the *H. nipponensis* HNIP-T2T assembly, we first mapped multi-platform sequencing data: MGI short reads, PacBio HiFi reads and ONT long reads achieved 99.72%, 99.97%, and 99.86% mapping rates (with 98.71–99.99% genome coverage; Fig. 5), confirming strong consistency with raw sequencing data. Transcriptome alignment also showed a higher unique mapping rate for HNIP-T2T than HNIP-V2 (Table 6), supporting superior structural accuracy for downstream analyses.

BUSCO (v5.8.0)[52] (Actinopterygii_odb10, 3,640 orthologs) benchmarking revealed 98.2% complete genes (97.5% single-copy) for HNIP-T2T—exceeding HNIP-V2's 96.7% (Table 4); protein-level BUSCO yielded 98.0% complete orthologs, validating structural integrity. Merqury[53] ($k = 19$) assigned a QV score of 35.11 (Table 4), consistent with T2T-level accuracy.

HNIP-T2T also outperformed HNIP-V2 in contiguity: its contig N50 (20.23 Mb) was 2.5 × longer, and it was gap-free (0 vs. 189 gaps in HNIP-V2; Table 4). Minimap2-derived[54] read coverage plots (Fig. 6) showed uniform depth across all 28 chromosomes, resolving HNIP-V2's fragmented coverage and gaps. Mummer[55] collinear alignment (Fig. 7) confirmed strict chromosome-level synteny between assemblies (93.01%/95.43% aligned bases for HNIP-T2T/HNIP-V2): diagonal high-similarity hits verified HNIP-T2T retained HNIP-V2's chromosomal framework while correcting local misassemblies (diagonal deviations in HNIP-V2 correspond to HNIP-T2T's linearity improvements). Presence-absence variation (PAV) analysis was performed using BWA (v0.7.17-r1188)[56] with the MEM algorithm (parameters: -w 500 -M -t 16; Table 7). The results revealed that the HNIP-T2T assembly exhibits a substantially expanded repertoire of PAVs—defined as sequences failing to align or showing <25% coverage—compared to HNIP-V2. Specifically, the PAV content increased from 1.75 Mb (0.34% of the genome) in HNIP-V2 to 7.55 Mb (1.43%) in HNIP-T2T. This expansion primarily reflects the successful filling of genomic gaps, while SNP rates remained conserved between the two versions (~0.21–0.22%). BUSCO assessment of gene set completeness (Actinopterygii_odb10) confirmed HNIP-T2T's superiority (3,569 complete orthologs), outperforming HNIP-V2 (3,485) and *H. transpacificus* (3,334; Fig. 8).

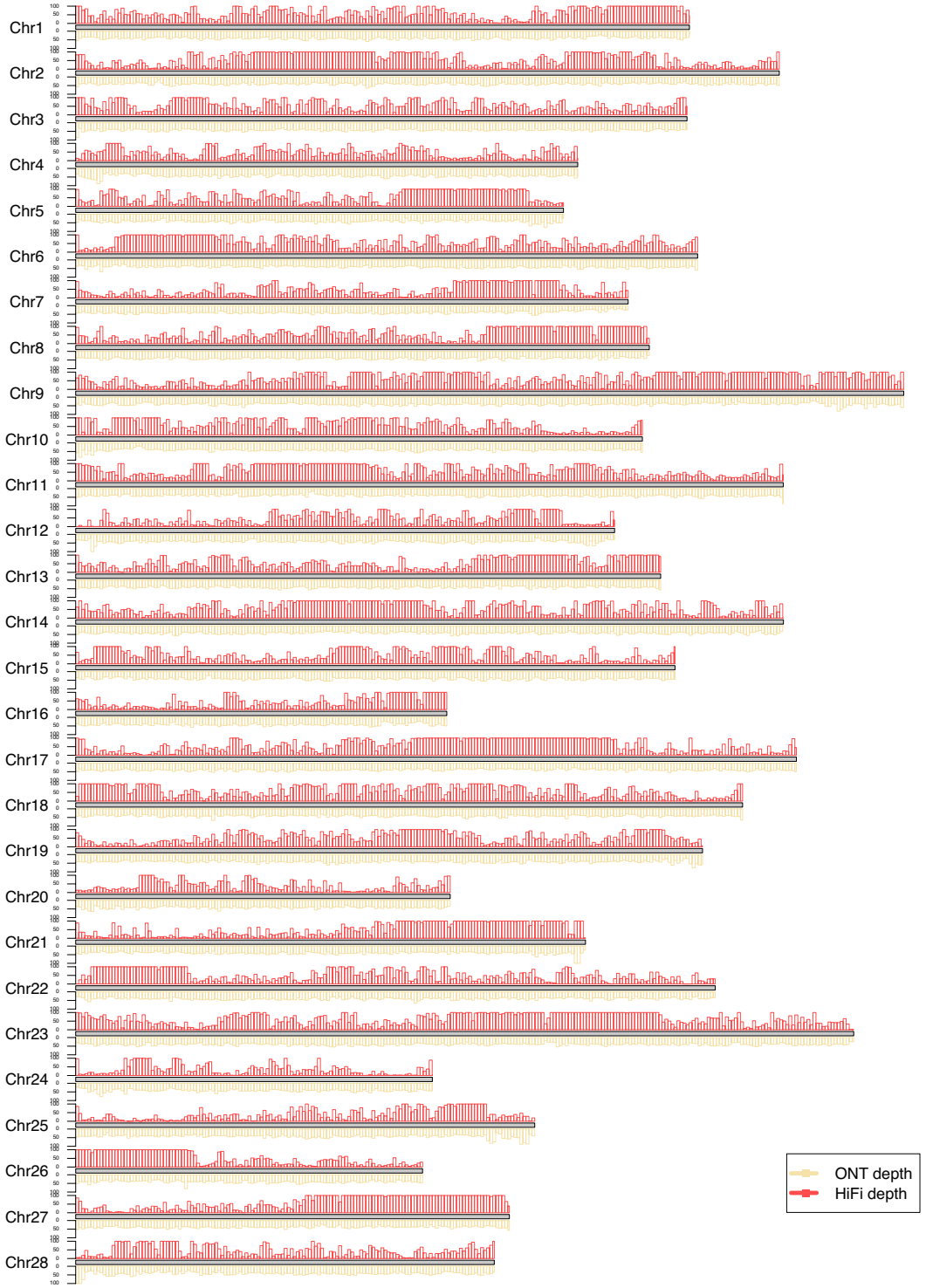

Fig. 6 The genome-read coverage plot (ONT and PacBio HiFi reads mapped via Minimap2).

Overall, HNIP-T2T represents a substantial improvement over HNIP-V2, with higher completeness, longer contiguity, gap-free structure, and robust mapping/transcriptome alignment performance.

Data availability

All data supporting this study have been publicly available. Raw sequencing data have been deposited in the NCBI Sequence Read Archive (SRA) database under the BioProject id PRJNA1282796[49], including RNA-seq data (SRR34259912 to SRR34259916), MGI genome survey data (SRR34259908), Hi-C reads (SRR34259911),

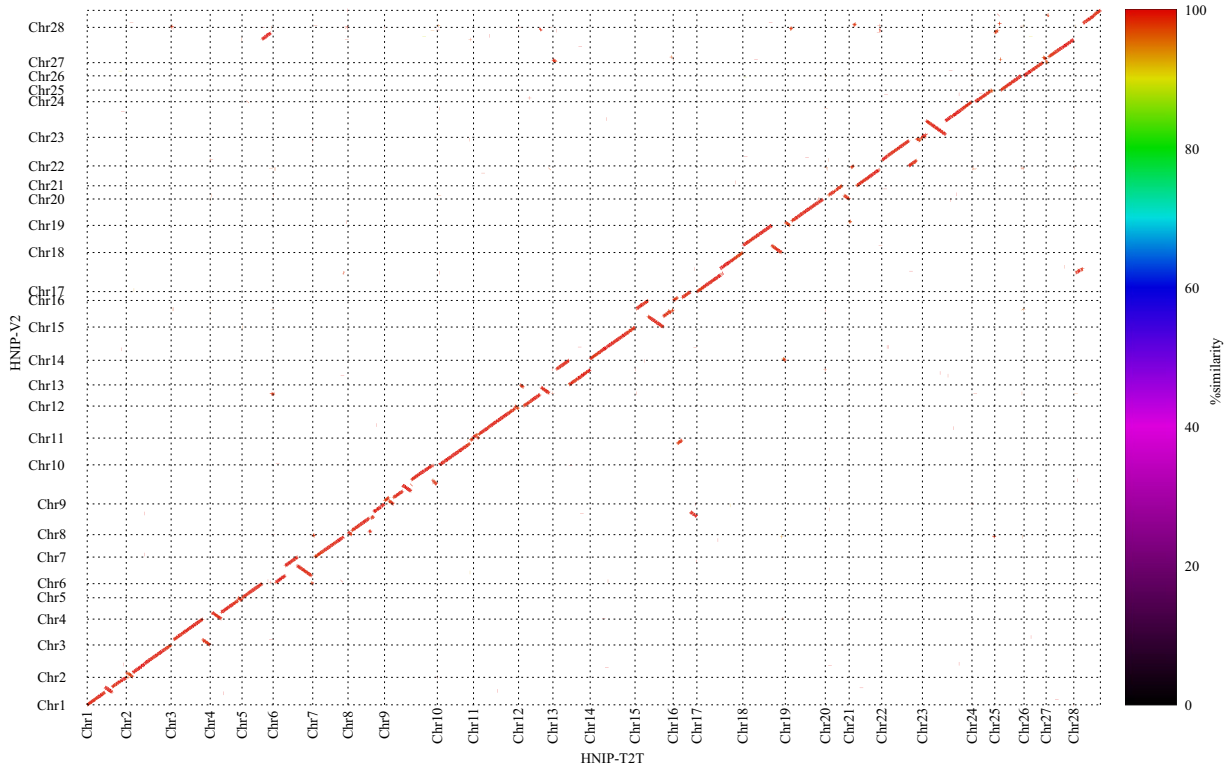

Fig. 7 Collinear alignment (Mummer) between HNIP-T2T and HNIP-V2 assembly of *H. nipponensis* genome.

Type	HNIP-T2T	HNIP-V2
Aligned Ratio /%	93.01	95.43
Breakpoints Number	55,017	54,999
Relocations Number	1,303	1,251
Translocations Number	3,008	3,170
Inversions Number	397	385
Insertions Number	13,693	16,209
InsertionSum	7.43	5.11
TandemIns Number	3,597	4,493
TandemInsSum /%	0.24	0.27
TotalSNPs /%	0.21	0.22
Presence-AbsenceVariations Number	3,001	1,257
Presence-AbsenceVariation Sum /%	1.43	0.34

Table 7. Summary of genome structure alignment data between **HNIP-T2T** and **HNIP-V2 of** *H. nipponensis* genome. Note: Aligned Ratio /%: Percentage of sequences aligned between compared genomes. Breakpoints Number: Count of genomic sequence disruption positions. Relocations Number: Number of genomic segments translocated to novel positions. Translocations Number: Count of segments transferred between non-homologous chromosomes. Inversions Number: Number of genomic segments with reversed orientation. Insertions Number: Count of inserted DNA fragments. InsertionSum: Total length or proportion of inserted sequences (unit as indicated in methods). TandemIns Number: Count of insertions with adjacent tandem repeat sequences. TandemInsSum /%: Proportion of genome occupied by tandem insertions. TotalSNPs /%: Genome-wide percentage of single-nucleotide polymorphisms. Presence-AbsenceVariations Number: Count of genomic regions present in one genome but absent in the other. Presence-AbsenceVariation Sum /%: Proportion of genome affected by presence-absence variations.

Nanopore long-read data (SRR34259910) and PacBio long-read data (SRR34259909). The genome assembly has been deposited at the NCBl GenBank under the accession number of GCA_054491055.1[50]. The genome assembly and gene structure annotation are also available on Figshare (https://doi.org/10.6084/m9.figshare.29672606.v1)[51].

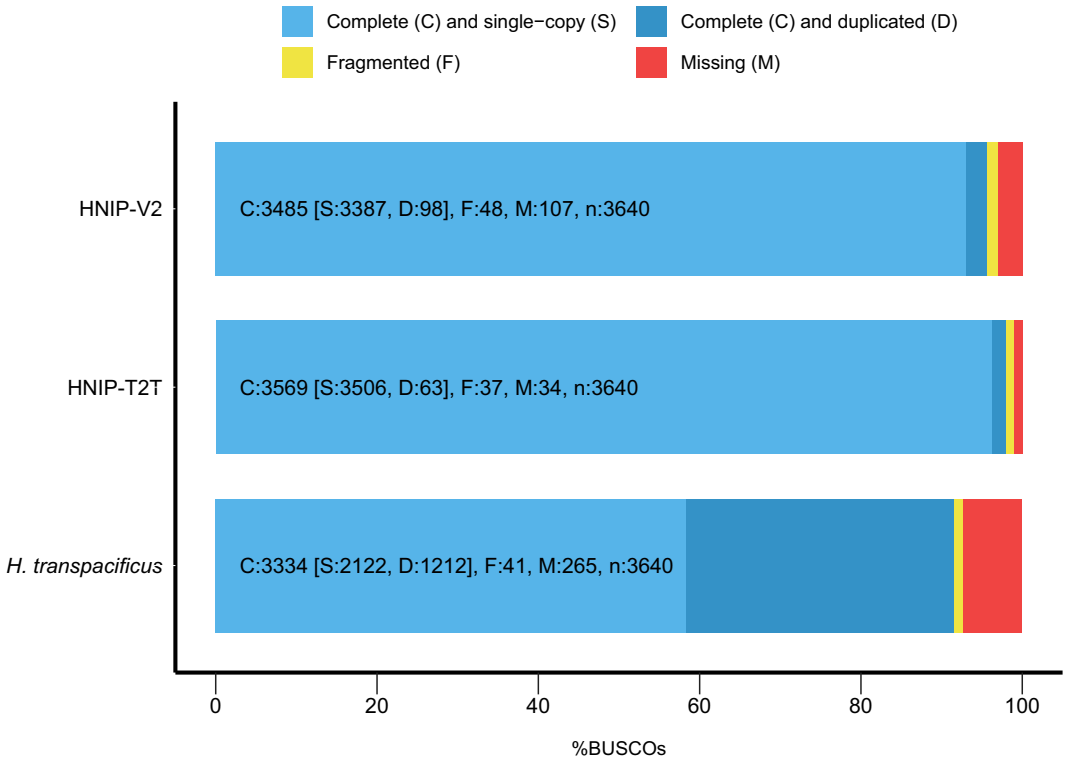

Fig. 8 BUSCO assessment of gene set completeness.

Code availability

All scripts and pipelines used for the genome assembly and gene annotation followed the standard manuals and protocols of the applied bioinformatics software. No specific code was developed for this study.

Received: 15 September 2025; Accepted: 12 March 2026;
Published online: 27 March 2026

References

1. Sakamoto, D. *et al.* Population size estimation of the pond smelt Hypomesus nipponensis in Lake Kasumigaura and Lake Kitaura, Japan. *Fisheries Science* **80**, 907–914, https://doi.org/10.1007/s12562-014-0791-1 (2014).
2. Xie, Y. *et al.* The fishes of genus Hypomesus and utilization of its resource (in Chinese) (Liaoning Science and Technology Press, 1992).
3. Yin, C., Chen, Y., Guo, L. & Ni, L. Fish Assemblage Shift after Japanese Smelt (Hypomesus nipponensis McAllister, 1963) Invasion in Lake Erhai, a Subtropical Plateau Lake in China. *Water* **13**, 1800, https://doi.org/10.3390/w13131800 (2021).
4. Choi, S. & Kim, E. B. Complete mitochondrial genome sequence and SNPs of the Korean smelt Hypomesus nipponensis (Osmeriformes, Osmeridae). *Mitochondrial DNA Part B* **4**, 1844–1845, https://doi.org/10.1080/23802359.2019.1613178 (2019).
5. Xuan, B. *et al.* Draft genome of the Korean smelt Hypomesus nipponensis and its transcriptomic responses to heat stress in the liver and muscle. G3 (Bethesda) **11**, https://doi.org/10.1093/g3journal/jkab147 (2021).
6. Zhu, C., Kuang, Y., Li, Z. & Tang, F. Chromosome-level draft genome assembly of Hypomesus nipponensis reveals transposable element expansion reshaping the genome structure. *Front Genet* **16**, 1502681, https://doi.org/10.3389/fgene.2025.1502681 (2025).
7. Shay, J. W. & Wright, W. E. Telomeres and telomerase: three decades of progress. *Nat Rev Genet* **20**, 299–309, https://doi.org/10.1038/s41576-019-0099-1 (2019).
8. Wu, M. *et al.* Segrosome assembly at the pliable parH centromere. *Nucleic Acids Res* **39**, 5082–5097, https://doi.org/10.1093/nar/gkr115 (2011).
9. Jain, M. *et al.* Linear assembly of a human centromere on the Y chromosome. *Nature Biotechnology* **36**, 321–323, https://doi.org/10.1038/nbt.4109 (2018).
10. Vollger, M. R. *et al.* Segmental duplications and their variation in a complete human genome. *Science* **376**, eabj6965, https://doi.org/10.1126/science.abj6965 (2022).
11. Yin, D. *et al.* Telomere-to-telomere gap-free genome assembly of the endangered Yangtze finless porpoise and East Asian finless porpoise. GigaScience 13, https://doi.org/10.1093/gigascience/giae067 (2024).
12. Zhou, Y. *et al.* Gap-free genome assembly of Salangid icefish Neosalanx taihuensis. *Scientific Data* **10**, 768, https://doi.org/10.1038/s41597-023-02677-z (2023).
13. Zhou, Y. *et al.* Telomere-to-telomere genome and resequencing of 231 individuals reveal evolution, genomic footprints in Asian icefish, Protosalanx chinensis. GigaScience 14, https://doi.org/10.1093/gigascience/giaf067 (2025).
14. Jiang, M. *et al.* The telomere-to-telomere gap-free reference genome and taxonomic reassessment of Siniperca roulei. GigaScience 14, https://doi.org/10.1093/gigascience/giaf068 (2025).
15. Cheng, H. *et al.* Efficient near telomere-to-telomere assembly of Nanopore Simplex reads. bioRxiv, https://doi.org/10.1101/2025.04.14.648685 (2025).
16. Healey, A., Furtado, A., Cooper, T. & Henry, R. J. Protocol: a simple method for extracting next-generation sequencing quality genomic DNA from recalcitrant plant species. *Plant Methods* **10**, 21, https://doi.org/10.1186/1746-4811-10-21 (2014).
17. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890, https://doi.org/10.1093/bioinformatics/bty560 (2018).

18. Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics* **13**, 278–289, https://doi.org/10.1016/j.gpb.2015.08.002 (2015).
19. Zhu, W. *et al*. Altered chromatin compaction and histone methylation drive non-additive gene expression in an interspecific Arabidopsis hybrid. *Genome Biology* **18**, 157, https://doi.org/10.1186/s13059-017-1281-4 (2017).
20. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770, https://doi.org/10.1093/bioinformatics/btr011 (2011).
21. Sun, H., Ding, J., Piednoël, M. & Schneeberger, K. findGSE: estimating genome size variation within human and Arabidopsis using k-mer frequencies. *Bioinformatics (Oxford, England)* **34**, 550–557, https://doi.org/10.1093/bioinformatics/btx637 (2018).
22. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods* **18**, 170–175, https://doi.org/10.1038/s41592-020-01056-5 (2021).
23. Hu, J. *et al*. NextDenovo: an efficient error correction and accurate assembly tool for noisy long reads. *Genome Biology* **25**, 107, https://doi.org/10.1186/s13059-024-03252-4 (2024).
24. Walker, B. J. *et al*. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963, https://doi.org/10.1371/journal.pone.0112963 (2014).
25. Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* **19**, 460, https://doi.org/10.1186/s12859-018-2485-7 (2018).
26. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357–359, https://doi.org/10.1038/nmeth.1923 (2012).
27. Servant, N. *et al*. HiC-Pro: An optimized and flexible pipeline for Hi-C data processing. Genome Biology 16, https://doi.org/10.1186/s13059-015-0831-x (2015).
28. Durand, N. *et al*. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems* **3**, 95–98, https://doi.org/10.1016/j.cels.2016.07.002 (2016).
29. Dudchenko, O. *et al*. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, eaal3327, https://doi.org/10.1126/science.aal3327 (2017).
30. Durand, N. C. *et al*. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell systems* **3**, 99–101, https://doi.org/10.1016/j.cels.2015.07.012 (2016).
31. Wang, G. & Yu, W. J. A preliminary study on the karyotype of Hypomesus olidus. Salmon Fishery 2(1), n.p. (in Chinese) (1989).
32. Xu, G. C. *et al*. LR_Gapcloser: a tiling path-based gap closer that uses long reads to complete genome assembly. Gigascience 8, https://doi.org/10.1093/gigascience/giy157 (2019).
33. Lin, Y. *et al*. quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. Hortic Res, https://doi.org/10.1093/hr/uhad127 (2023).
34. Jurka, J. *et al*. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110**, 462–467, https://doi.org/10.1159/000084979 (2005).
35. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics (Oxford, England)* **21**(Suppl 1), i351–358, https://doi.org/10.1093/bioinformatics/bti1018 (2005).
36. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic acids research* **35**, W265–268, https://doi.org/10.1093/nar/gkm286 (2007).
37. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics Chapter* **4**, 4.10.11–14.10.14, https://doi.org/10.1002/0471250953.bi0410s25 (2009).
38. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**, 573–580, https://doi.org/10.1093/nar/27.2.573 (1999).
39. Liu, L. *et al*. Multiomics analysis reveals signatures of selection and loci associated with complex traits in pigs. *Imeta* **3**, e250, https://doi.org/10.1002/imt2.250 (2024).
40. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods* **12**, 357–360, https://doi.org/10.1038/nmeth.3317 (2015).
41. Kovaka, S. *et al*. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biology* **20**, 278, https://doi.org/10.1186/s13059-019-1910-1 (2019).
42. Keilwagen, J., Hartung, F. & Grau, J. GeMoMa: Homology-Based Gene Prediction Utilizing Intron Position Conservation and RNA-seq Data. *Methods Mol Biol* **1962**, 161–177, https://doi.org/10.1007/978-1-4939-9173-0_9 (2019).
43. Haas, B. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research* **31**, 5654–5666, https://doi.org/10.1093/nar/gkg770 (2003).
44. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Research* **27**, 49–54, https://doi.org/10.1093/nar/27.1.49 (1999).
45. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28**, 27–30, https://doi.org/10.1093/nar/28.1.27 (2000).
46. Tatusov, R., Galperin, M., Natale, D. & Koonin, E. The COG Database: A Tool for Genome-Scale Analysis of Protein Functions and Evolution. Nucleic Acids Research 28, https://doi.org/10.1093/nar/28.1.33 (2000).
47. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* **12**, 59–60, https://doi.org/10.1038/nmeth.3176 (2015).
48. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240, https://doi.org/10.1093/bioinformatics/btu031 (2014).
49. *NCBI Sequence Read Archive*. https://identifiers.org/ncbi/insdc.sra:SRP595455 (2025).
50. *NCBI GenBank*. https://identifiers.org/ncbi/insdc.gca:GCA_054491055.1 (2026).
51. Zhou, Y. Telomere-to-telomere genome assembly of Hypomesus nipponensis. *figshare. Dataset*. https://doi.org/10.6084/m9.figshare.29672606.v1 (2025).
52. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212, https://doi.org/10.1093/bioinformatics/btv351 (2015).
53. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol* **21**, 245, https://doi.org/10.1186/s13059-020-02134-9 (2020).
54. Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572–4574, https://doi.org/10.1093/bioinformatics/btab705 (2021).
55. Marçais, G. *et al*. MUMmer4: A fast and versatile genome alignment system. *PLoS Comput Biol* **14**, e1005944, https://doi.org/10.1371/journal.pcbi.1005944 (2018).
56. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv: Genomics* https://doi.org/10.48550/arXiv.1303.3997 (2013).

Acknowledgements

This work was financially supported by the Earmarked Fund for the National Key R&D Program of China (Grant No. 2023YFD2400900) and the Modern Agricultural Technology System Grant (CARS-46).

Author contributions

D. Xu designed and conceived the study. Y. Zhou, D. Fang, Y. You and X. Li collected the samples, conducted experiments. F. Tang, Y. Bai and M. Zhang performed bioinformatics analysis. Y. Zhou, G. Deng and D. Xu wrote and revised the manuscript. All authors have 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 Y.Z. or D.X.

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.