scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Near telomere-to-telomere genome assembly of the fourfinger threadfin (*Eleutheronema tetradactylum*)

Huijuan Zhang[1], Yifei Pan[1], Benxun Miao[1], Linjuan Wang[1], Anna Zheng[1], Minxuan Jin[1], Jiandong Zhang[1], Baogui Tang[1], Bei Wang[1,2], Jiansheng Huang[1], Jing Li[1], Dee-hwa Chong[3] & Zhongliang Wang[1,2,4] ✉

The fourfinger threadfin (*Eleutheronema tetradactylum*) is a euryhaline fish distributed across the Indo-West Pacific, from the Persian Gulf to Australia. However, the lack of high-quality genomic resources has limited genomic-level studies of its evolution, conservation, and aquaculture. Here, we present the first Near telomere-to-telomere (T2T) genome assembly of *E. tetradactylum*, generated using Pacific Biosciences (PacBio) High-Fidelity (HiFi) sequencing, Oxford Nanopore Technologies (ONT) ultra-long reads, and Hi-C chromatin conformation capture. The final assembly spans 585.38 Mb across 83 contigs, with a contig N50 of 22.14 Mb and 98.76% of sequences anchored to 26 chromosomes. We annotated 22,362 protein-coding genes and identified that repetitive sequences constitute 18.09% of the genome. This high-quality T2T assembly demonstrates significant improvements in contiguity and completeness over previously available genomes, providing an invaluable resource to accelerate genetic research, advance molecular breeding, and inform conservation strategies for *E. tetradactylum*.

Background & Summary

The fourfinger threadfin (*Eleutheronema tetradactylum*), a member of the family Polynemidae[1], is a euryhaline pelagic fish inhabiting offshore waters of the Indo-West Pacific, spanning from the Persian Gulf to Australia[2,3]. Distinguished by its elongated, flattened body and four filamentous pectoral fins (Fig. 1), this species employs these fins for sensory detection of prey. As a commercially valuable species, *E. tetradactylum* is highly regarded in aquaculture for its rapid growth rate and high-quality meat[4]. However, overfishing and habitat degradation have led to significant population declines, resulting in its classification as Near Threatened on the International Union for Conservation of Nature (IUCN) Red List in 2014[5]. Despite its importance, research has been constrained by a lack of high-quality genomic data, although studies have explored its population structure[6,7], disease control[8,9], farming techniques[10], reproduction, basic biology[11,12], and responses to environmental stressors[2,13].

Although chromosome-level genome assemblies of E. tetradactylum have been published[14], they are constrained by gaps, structural discontinuities, and incomplete annotations. Recent advancements in long-read sequencing technologies, such as PacBio and ONT, combined with improved assembly algorithms, have enabled the production of gap-free telomere-to-telomere (T2T) genomes. These assemblies resolve fragmented regions, enhance chromosomal continuity, and facilitate comprehensive variant detection[15,16].

In this study, we generated a Near T2T genome assembly for *E. tetradactylum* using PacBio HiFi reads, ONT ultra-long reads, and Hi-C data. The assembly totals 585.38 Mb, comprising 83 contigs with an N50 of 22.14 Mb, and anchors 98.76% of sequences to 26 chromosomes. We annotated 22,362 protein-coding genes, achieving BUSCO completeness scores of 99.53% for the genome and 99.04% for annotations. This resource surpasses

[1]College of Fisheries, Guangdong Ocean University, Zhanjiang, China. [2]Guangdong Provincial Key Laboratory of Aquatic Animal Disease Control and Healthy Culture, Zhanjiang, China. [3]Ichthyological Society of Hong Kong, Hong Kong, China. [4]Guangdong Provincial Marine Fish Technology Innovation Center, Zhanjiang, China. ✉e-mail: wangzl@gdou.edu.cn

Fig. 1 The map of *Eleutheronema tetradactylum*.

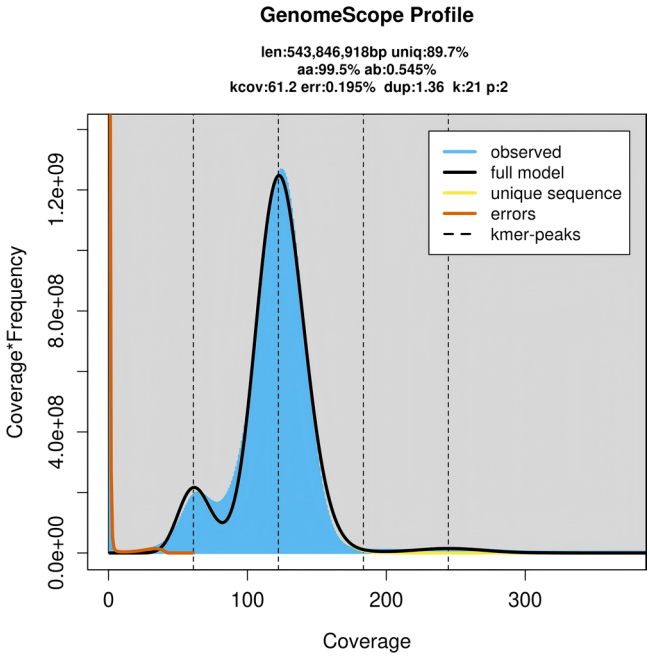

Fig. 2 Overview of the 21-mer frequency distribution in the *E. tetradactylum* genome. The x-axis indicates the coverage of the K-mer, the y-axis represents the k-mer frequency for a given depth.

Sequencing technology	Bases (Gb)	Reads number	Mean read length (bp)	GC Content (%)
HiFi	32.77	2,097,480	15,623.87	40.02
ONT	31.74	1,022,583	31,040	40.30
Hi-C	139.39	929,374,606	150	40.21
RNA-seq	21.94	147,098,490	150	47.71

Table 1. Sequencing data for the *E. tetradactylum* genome assembly.

prior assemblies in contiguity and completeness, providing a foundational tool for advancing genetic, evolutionary, and conservation research on *E. tetradactylum*.

Methods

Sample collection. A two-year-old male *E. tetradactylum*, sourced from a local fishery farm in Zhanjiang, Guangdong Province, China, was used for this study. Tissues including muscle, eye, brain, liver, heart, spleen, kidney, and gill, were collected for genomic and transcriptomic sequencing. All samples were flash-frozen in liquid nitrogen and stored at −80 °C. was All procedures were approved by the Institutional Review Board on Bioethics and Biosafety of BGI-Shenzhen, China (No. FT18134).

Library construction and sequencing. Genomic DNA was extracted from muscle tissue for PacBio Single Molecule Real-Time (SMRT), Hi-C, Oxford Nanopore Technologies (ONT), and short-read genome survey sequencing. For SMRT sequencing, high-quality DNA was used to construct genomic libraries according to PacBio's standard protocol (Pacific Biosciences, CA, USA) and sequenced on the PacBio Sequel II platform (Pacific Biosciences, Menlo Park, CA, USA) in Circular Consensus Sequence (CCS) mode. The raw data was filtered to obtain high-precision HiFi reads.

Peak	Kmers [number]	Kmers [proportion]	Summit B/(A + B)	Summit A + B
AB	18,358,967	0.57	0.49	129.92
AABB	13,796,282	0.43	0.49	246.74

Table 2. Smudgeplot analysis statistics for ploidy determination.

Genome assembly statistics	New genome	Published genome (PRJNA975807)
Total length (Mb)	585.38	582.0
Number of scaffolds	77	26
N50 length (scafold) (Mb)	23.88	23.87
N90 length (scaffold) (Mb)	17.41	—
Number of contig	83	80
N50 length (contig) (Mb)	22.14	18.26
N90 length (contig) (Mb)	12.76	—
Number of chromosomes	26	26
Anchoring rate (%)	98.76	98.84
GC Content (%)	40.2	40.45
BUSCO	3,623 (99.53%)	2,550 (98.6%)
Gene prediction	22,362	21667

Table 3. Statistics of the *E. tetradactylum* genome assembly and comparison with a prior assembly.

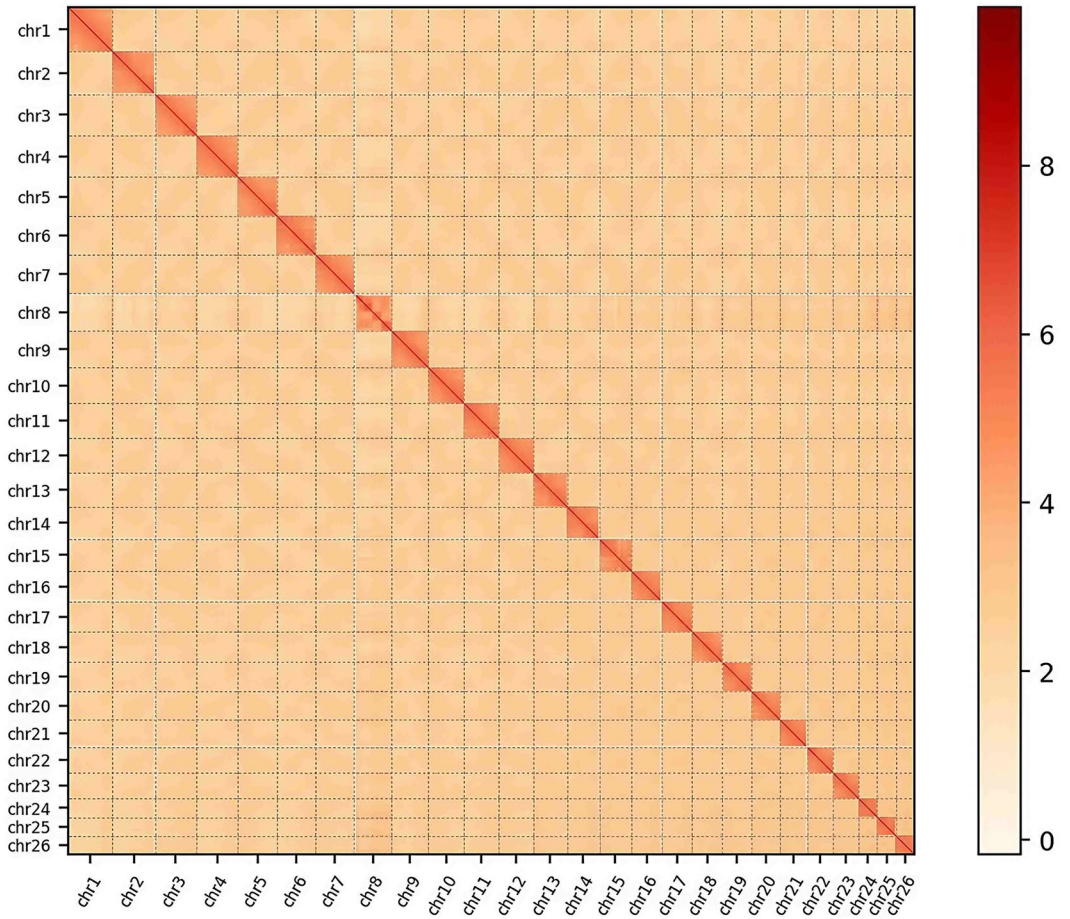

Fig. 3 Hi-C interaction heatmap of the *E. tetradactylum* genome. The x- and y-axes represent genomic positions (N*bin). Color intensity (yellow: low; red: high) indicates interaction strength. The first 26 squares represent the 26 chromosomes, followed by unanchored sequences.

For ONT sequencing, an ultra-long library was constructed and sequenced on a PromethION flow cell (Oxford Nanopore Technologies Co., UK). Raw reads were filtered for quality (QV $\geq$ 7) using base-calling

Name	Length (Mb)	Number of gaps	Number of telomeres
chr1	30.24	0	2
chr2	29.00	0	1
chr3	28.83	0	1
chr4	28.27	0	2
chr5	26.73	0	2
chr6	26.49	0	2
chr7	26.43	0	2
chr8	25.63	3	1
chr9	25.00	0	2
chr10	24.23	0	2
chr11	23.88	1	2
chr12	23.88	0	2
chr13	22.81	0	1
chr14	22.15	0	1
chr15	21.75	0	1
chr16	21.03	0	1
chr17	20.71	0	1
chr18	20.70	0	1
chr19	20.08	0	1
chr20	19.03	0	1
chr21	18.72	0	1
chr22	17.90	0	1
chr23	17.42	0	1
chr24	12.81	0	1
chr25	12.83	1	1
chr26	12.76	0	2

Table 4. Assembly statistics of chromosomes.

software, adapters were removed with Porechop (https://github.com/rrwick/Porechop), and reads shorter than 30 kb or with mean quality <90% were discarded using Filtlong (https://github.com/rrwick/Filtlong).

Hi-C libraries were constructed from muscle tissue following established protocols[17]. Tissue was cross-linked with formaldehyde, digested with a restriction enzyme, biotin-labeled, and ligated. After reversing cross-links and purifying DNA, fragments were sheared to ~300 bp, and paired-end libraries were sequenced on the DNBSEQ platform.

Total RNA was extracted from eight tissues (eye, brain, liver, heart, spleen, kidney, muscle, gill) using TRIzol reagent (Invitrogen). Paired-end sequencing was performed on the MGI-SEQ. 2000 platform.

Sequencing generated 32.77 Gb of HiFi data, 31.74 Gb of ONT data, 139.39 Gb of Hi-C data, and 21.94 Gb of RNA-seq data (Table 1).

Genome survey and assembly. For the genome survey, DNA libraries with 300–400 bp inserts were constructed. Then, DNA was purified, quantified, and sequenced from both ends using the DNBSEQ platform to obtain raw reads. Quality filtering of raw reads was performed using Fastp (v0.23.2; parameters: default)[18], and K-mer frequency (K = 21) was calculated with Jellyfish (v2.3.0; parameters: -m 21 -s 1000000000)[19]. Based on K-mer distribution, GenomeScope 2.0 (v2.0; parameters: -k 21 -p 2)[20] estimated the genome size to be 543.84 Mb, with a peak 21-mer depth of 120 (Fig. 2). The heterozygosity and repeat rates were found to be 0.545% and 10.328%, respectively. Smudgeplot (v0.2.3dev; parameters: -k21 -m100 -ci1 -cs1000)[20] determined the species' ploidy as AB type, indicating diploidy (Table 2).

The draft assembly of *E. tetradactylum* was performed using HiFi data combined with ONT ultra-long reads and Hi-C reads. The assembly was carried out with HiFiasm (v0.19.6; parameters: default)[21], followed by redundancy removal with Purge Haplotigs (v1.0.4; parameters: default)[22]. This high-quality genome assembly served as the foundation for subsequent construction of chromosomes using the Hi-C reads.

Hi-C reads were aligned to the draft[23], and the 3D-DNA pipeline, which included splitting, anchoring, sorting, orienting, and merging contigs or scaffolds, was employed to achieve chromosome-level scaffolding[24]. An interaction matrix was generated with Juicer (v1.5; parameters: chr_num 24)[25] and manually refined using Juicebox (v1.11.08; parameters: default)[26].

Ultra-long Oxford Nanopore Technologies (ONT) reads were aligned to chromosomes using minimap2 (v2-2.24; parameters: ont: -ax map-ont ccs: -ax map-hifi)[27] to generate consensus sequences. These consensus sequences were then aligned to the ends of the chromosomes using blastn (v2.11.0+; parameters: -outfmt 7), and sequences with coverage ≥90% were used to replace the telomere sequences on the chromosomes based on their alignment positions. Gaps between contigs were filled using TGS-GapCloser (v1.2.0; parameters:–min_nread 10)[28] by leveraging the coverage information between ultra-long ONT reads and the assembled

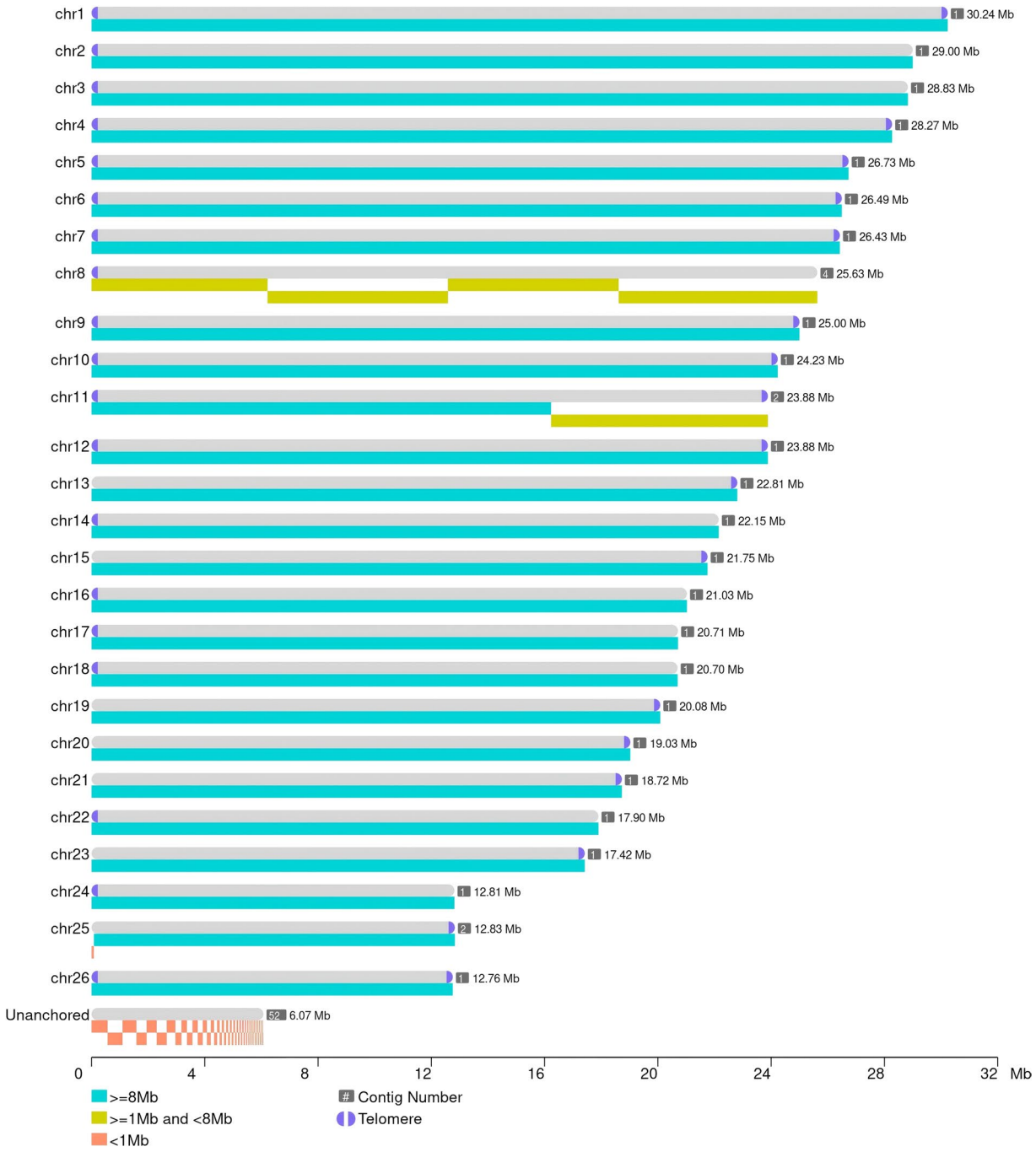

Fig. 4 An overview of the T2T gap-free reference genome of *E. tetradactylum*.

contigs to perform contig extension. Subsequent polishing was carried out with Pilon (v1.23; parameters:–fix all–changes)[29] using short-read sequencing data to correct errors in the extended and gap-filled genome, yielding the final telomere-to-telomere assembly of *E. tetradactylum*.

The final assembly spans 585.38 Mb across 77 scaffolds (26 chromosomes), with scaffold N50 of 23.88 Mb, contig N50 of 22.14 Mb, and 98.76% anchoring rate (Table 3). A Hi-C interaction heatmap confirmed high-quality chromosome assignments (Fig. 3). A total of 36 telomeric sequences were identified at the ends of the 26 chromosomes by searching the entire genome for the telomeric repeat motif (TTAGGG) (Table 4). The genomic positions of these telomeres and their distribution across contigs were annotated and visualized (Fig. 4).

Repeats annotation. Repetitive elements were identified using a combination of *de novo* and homology-based approaches. Tandem repeats were predicted with TRF (v4.09; default)[30]. Homology searches employed RepeatMasker (v4.0.9; default)[31] against the RepBase library (http://www.girin-st.org/repbase).

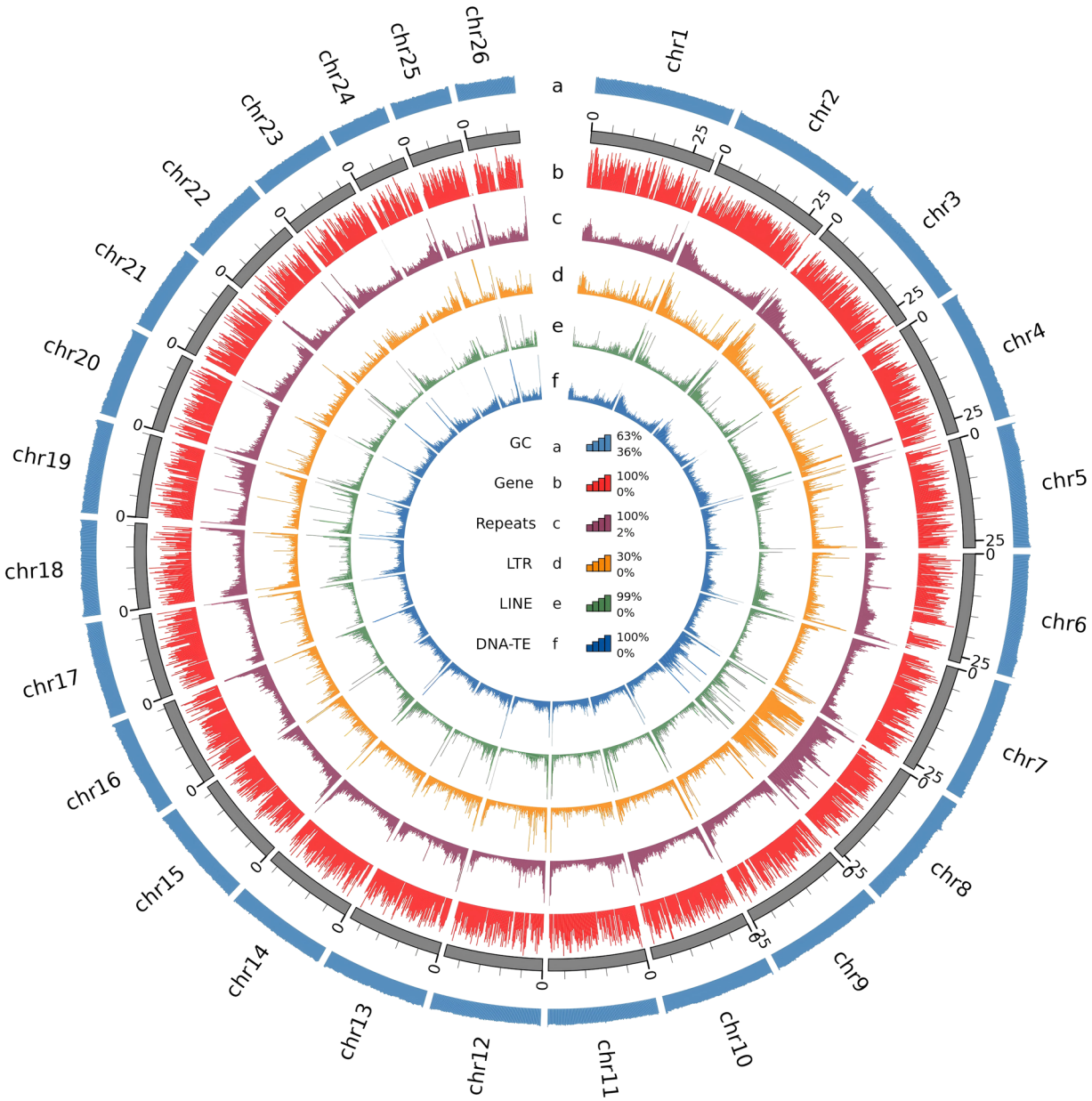

Fig. 5 Genomic landscape of the *E. tetradactylum* chromosome-level assembly. Metrics were calculated using a window size of approximately 200 kb. Circos plot from the outer to the inner layers represents the following: (a) GC content (range: 36%–56%); (b) gene density (range: 0%–100%); (c) repeat density (range: 0%–100%); (d) LTR retroelement density (range: 0%–24%); (e) LINE density (range: 0%–61%); and (f) DNA transposon density (range: 0%–88%).

Additionally, RepeatModeler (open-4.0.9; parameters: default)[32] and LTR_FINDER_parallel (v1.0.7; parameters: default)[33] were used to construct a *de novo* repeat library for *E. tetradactylum*, followed by a further *de novo* prediction using RepeatModeler. By integrating results from TRF, RepeatMasker, RepeatProteinMask, and *de novo* methods, and subsequently removing redundancies, we determined that repeat sequences and transposable elements (TEs) constitute approximately 18.09% and 16.69% of the *E. tetradactylum* genome, respectively (Fig. 5). Of which, repetitive DNAs, LINEs, SINEs and LTRs covered 8.69%, 3.19%, 0.29% and 1.70% of the entire genome, respectively (Table 5). This repeat content is comparable to that in *Lates calcarifer* (18.6%)[34] but higher than in *oreochromis niloticus* (14%)[35].

Gene prediction and functional annotation. To annotate genes in the *E. tetradactylum* genome, we conducted both structural and functional annotation. Gene structure annotation aimed to predict gene positions and structures through homology-based and *de novo* approaches, while functional annotation determined the biological roles and metabolic pathways associated with predicted gene products.

Type	RepBase TEs		TE Proteins		*De novo*		Combined TEs	
	Length (bp)	% in Genome	Length (bp)	% in Genome	Length (bp)	% in Genome	Length (bp)	% in Genome
DNA	34,601,389	5.91	3,713,038	0.63	21,899,652	3.74	50,892,902	8.69
LINE	12,531,239	2.14	7,044,708	1.20	10,027,160	1.71	18,668,401	3.19
SINE	1,271,331	0.22	0	0.00	606,641	0.10	1,704,600	0.29
LTR	7,070,835	1.21	1,522,057	0.26	2,596,525	0.44	9,924,602	1.70
Satellite	3,782,983	0.65	0	0.00	97,452	0.02	3,872,398	0.66
Simple_repeat	0	0.00	0	0.00	425,178	0.07	425,178	0.07
Other	5,685	0.00	150	0.00	0	0.00	5,835	0.00
Unknown	313,473	0.05	11,007	0.00	20,209,818	3.45	20,487,336	3.50
Total	52,829,496	9.02	12,280,039	2.10	54,903,217	9.38	97,672,199	16.69

Table 5. Statistics of repeat sequence classification in the *E. tetradactylum* genome.

Gene set	Number	Average gene length (bp)	Average CDS length (bp)	Average exon per gene	Average exon length (bp)	Average intron length (bp)
denovo/Genscan	29645	13745.88	1602.27	9.02	177.69	1514.67
denovo/AUGUSTUS	25452	10476.18	1503.15	8.74	172.00	1159.42
homo/*P. olivaceus*	53936	21835.29	2451.63	14.18	172.89	1470.66
homo/*O. latipes*	45401	19808.97	2358.65	13.24	178.16	1425.77
homo/*C. gibelio*	92148	20335.46	2199.32	11.87	185.29	1668.55
homo/*D. rerio*	90462	12801.83	2438.58	13.32	183.13	1756.34
homo/*O. latipes*	12321	10252.33	1079.83	4.56	236.86	2577.35
trans.orf/RNA-seq	16372	17763.22	1971.63	12.51	331.57	1182.96
BUSCO	3644	8747.95	1656.21	10.42	159.00	753.11
MAKER	22261	15938.62	1676.26	10.94	324.93	1246.18
HiFAP	22362	14620.72	1811.74	10.80	270.42	1193.41

Table 6. Statistics of protein-coding gene predictions in the *E. tetradactylum* genome.

For gene structure annotation, we combined three strategies, including homology-based predictions, *de novo* prediction and RNA-sequencing-assisted prediction. we utilized Exonerate (v2.2.0; parameters: model protein2genome)[36] and Liftoff (v1.6.3; parameters: showtargetgff 1)[37] to align *E. tetradactylum* genome sequence with protein sequences from closely related species (*Paralichthys olivaceus, Oryzias latipes, Carassius gibelio, Danio rerio*, and *Oryzias latipes*) for homology-based prediction. *De novo* predictions were performed using AUGUSTUS (v3.3.2; parameters: default)[38] and Genscan (v1.0; parameters: default)[39]. Additionally, RNA-seq data were mapped onto the *E. tetradactylum* genome, with transcripts and protein-coding genes predicted separately using StringTie (v1.3.5; parameters: default)[40] and TransDecoder (v5.5.0; https://github.com/TransDecoder/TransDecoder) with default parameters. The predictions from these methods were integrated into a high-quality, non-redundant gene set using MAKER 2 (v2.31.10; parameters: default)[41].

For gene function annotation, we compared the protein sequences of the genome derived from structure annotation against various databases, including GO[42], KEGG[43], Swissprot[44], TrEMBL[45], NR[46], KOG (https://ftp.ncbi.nih.gov/pub/COG/KOG/) and AnimalTFDB[47]. This analysis, conducted using diamond (v2.0.14; parameters:–evalue 1e-05)[48] software, provided insight into protein functions, metabolic pathways, and additional characteristics. To further identify conserved sequences, motifs, and structural domains, we analyzed Pfam[49] and InterPro[50] databases by using InterProScan (v5.61–93.0; parameters:–seqtype p–formats TSV–gote rms –pathways -dp)[51]. Pathway annotation was performed using KOBAS (v3.0; parameters: -t blastout: tab-sko)[52] against the KEGG database Table 5.

Overall, we predicted 22362 protein-coding genes, with average gene length of 14620.72 bp, CDS length of 1811.74 bp,10.80 exons per gene, and exon length of 270.42 bp (Table 6). And then we predicted a total of 22,046 genes (98.59% of the total predicted genes) and 37,591 mRNA (98.71%) of the total predicted transcript) were successfully annotated (Fig. 6 and Table 7).

Non-coding RNAs were predicted using BLASTN(v2.11.0+; parameters: -evalue 1e-5)[53] for rRNAs, tRNAscan-SE (v1.3.1; parameters: default)[54] for tRNAs, and Infernal (v1.3.3) against Rfam (v14.8; parameters: cmscan --rfam --nohmmonly)[55] for miRNAs and snRNAs. We identified 791 miRNAs, 1594 tRNAs, 1102 rRNAs, and 651 snRNA (Table 8).

Genome collinearity analysis. To investigate the conservation of genome structure, a synteny analysis was performed between the coding genes of *E. tetradactylum* and a related species, *E. rhadinum*, using JCVI (v1.1.22; parameters: "jcvi.compara.catalog ortholog --dbtype = prot --cscore 0.99 jcvi.compara.synteny screen --minspan = 70–align-chromosomes")[56]. Both species share a 2n = 52 karyotype and exhibit high collinearity, indicating conserved synteny (Fig. 7).

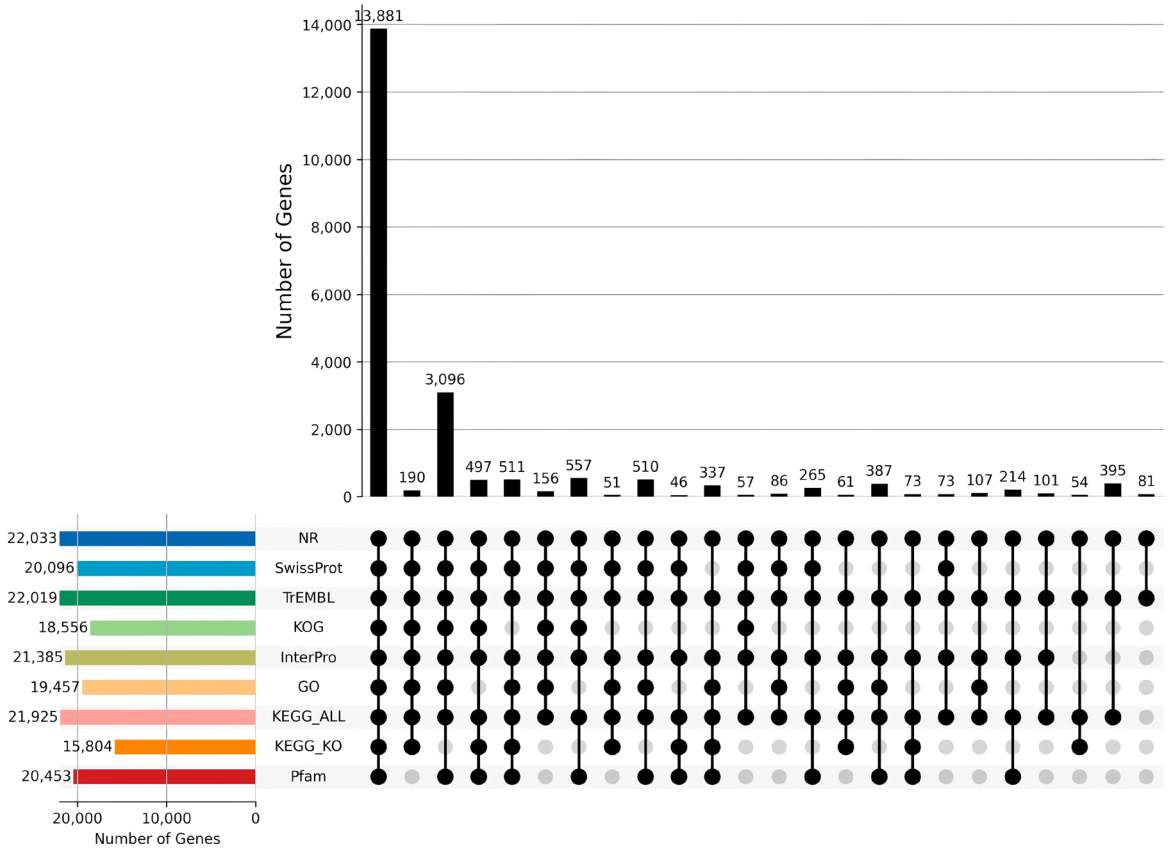

Fig. 6 UpSet plot of gene functional annotations across nine databases: NR, SwissProt, TrEMBL, KOG, InterPro, GO, KEGG-ALL, KEGG-KO, and Pfam.

Type	Gene		mRNA	
	Number	Percent (%)	Number	Percent (%)
Total	22,362	100	38,083	100
Annotated	22,046	98.59	37,591	98.71
NR	22,033	98.53	37,566	98.64
SwissProt	20,096	89.87	34,205	89.82
TrEMBL	22,019	98.47	37,543	98.58
KOG	18,556	82.98	31,740	83.34
TF	5,345	23.90	9,341	24.53
InterPro	21,385	95.63	36,047	94.65
GO	19,457	87.01	32,769	86.05
KEGG_ALL	21,925	98.05	37,417	98.25
KEGG_KO	15,804	70.67	27,356	71.83
Pfam	20,453	91.46	33,982	89.23
Unannotated	316	1.41	492	1.29

Table 7. Functional annotation of protein-coding genes in the *E. tetradactylum* genome. Note: Nine databases (Nr, SwissProt, TrEMBL, KOG, TF, InterPro, GO, KEGG, Pfam) were used for functional annotation.

Type	Copy	Average length (bp)	Total length (bp)	% of genome
miRNA	791	86	68,001	0.011616
tRNA	1,594	76	120,722	0.020623
rRNA	1,102	179	197,729	0.033778
snRNA	651	149	97,069	0.016582

Table 8. Statistics of non-coding RNAs in the *E. tetradactylum* genome.

BUSCO genome completeness score	Genome	Annotation
Complete BUSCOs (C)	3,623 (99.53%)	3,605 (99.04%)
Complete and single-copy BUSCOs (S)	3,614 (99.29%)	3,061 (84.09%)
Complete and duplicated BUSCOs (D)	9 (0.25%)	544 (14.95%)
Fragmented BUSCOs (F)	16 (0.44%)	12 (0.33%)
Missing BUSCOs (M)	1 (0.03%)	23 (0.63%)
Total BUSCO groups searched	3,640 (100%)	3,640 (100%)

Table 9. BUSCO completeness and accuracy evaluation of the *E. tetradactylum* genome and annotations.

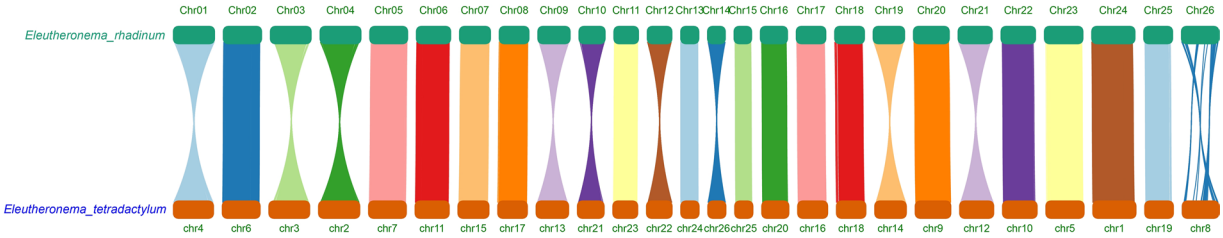

Fig. 7 Synteny analysis between *E. tetradactylum* and *E. rhadinum* genomes.

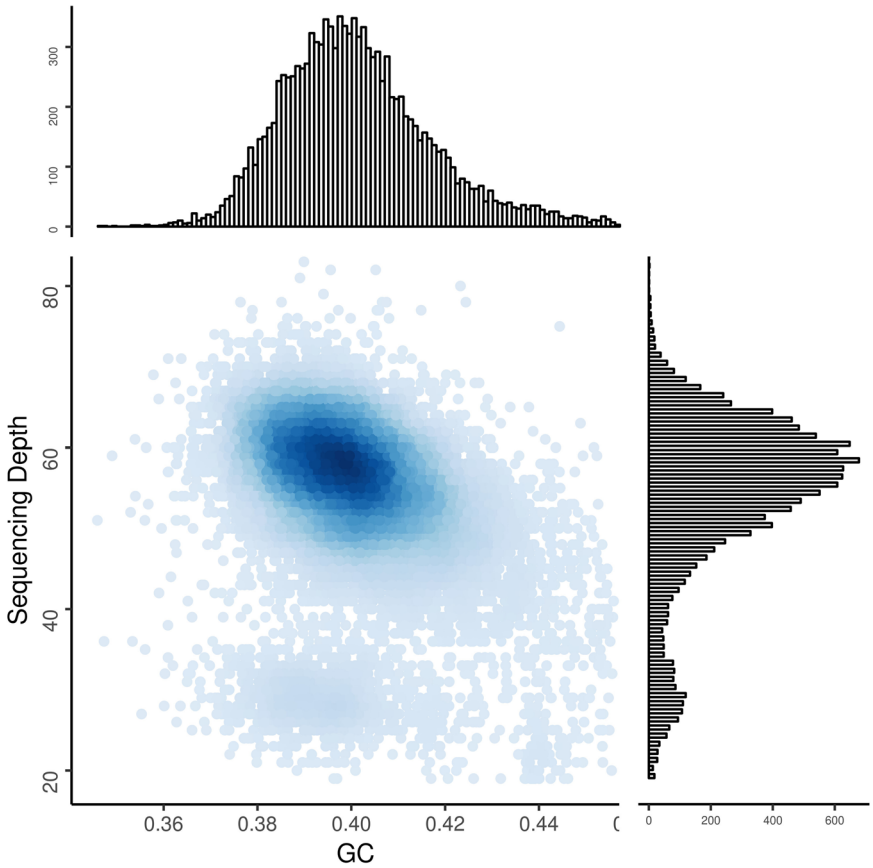

Fig. 8 GC content and sequencing depth distribution. The x-axis represents the GC content; the y-axis represents the average depth.

Data Records

The final telomere-to-telomere genome assembly for *E. tetradactylum* have been deposited in the National Center for Biotechnology Information (NCBI) GenBank database under accession number JBIEKL000000000[57]. Annotated coding sequences and protein sequences have been submitted to Figshare (https://doi.org/10.6084/m9.figshare.30164734)[58]. Raw sequencing reads (HiFi, Hi-C, ONT, genome survey, and RNA-seq) are deposited in the NCBI Sequence Read Archive (SRA) under accession number SRP538810[59]. All data are publicly accessible without restriction.

Technical Validation

Genome assembly and gene annotation quality assessment. Assembly and annotation completeness were evaluated with BUSCO (v5. 4. 3; parameters: default)[60] against the actinopterygii_odb10 lineage. The genome recovered 99.53% BUSCOs (99.29% single-copy, 0.25% duplicated, 0.44% fragmented, 0.03% missing; Table 9). Annotations recovered 99.04% (84.09% single-copy, 14.95% duplicated, 0.33% fragmented, 0.63% missing; Table 9).

The PacBio HiFi reads were aligned to the assembly using minimap2 (v2.12, parameters: -ax map-pb)[27], achieving 99.72% mapping and 99.84% coverage. GC content and depth were uniform across 100-kb windows (Fig. 8). Short reads were aligned with samtools (v1.17, parameters: sort -m 1 G)[27], picard (v2.25.6; https://broadinstitute.github.io/picard/) and, GATK(v4.4.0.0; https://broadinstitute.github.io/picard/), revealing heterozygous SNP and InDel rates of 0.279% and 0.111%, with no homozygosity.

In this study, we successfully achieved a T2T assembly for ten chromosomes: Chr1, Chr4, Chr5, Chr6, Chr7, Chr9, Chr10, Chr11, Chr12, and Chr26. For the remaining chromosomes, telomeres were identified at only one terminus. The difficulty in achieving complete T2T status for these sequences is likely attributable to the presence of recalcitrant genomic regions characterized by high complexity and extreme repetitive content (Fig. 5). Despite the utilization of current ONT ultra-long reads, the structural intricacy of these regions remains challenging to fully resolve. We anticipate that future advancements in sequencing read lengths and the continuous refinement of T2T assembly algorithms will eventually overcome these limitations, enabling the complete, gap-free assembly of the entire *E. tetradactylum* genome.

Data availability

The final telomere-to-telomere genome assembly for *E. tetradactylum* is available under GenBank accession JBIEKL000000000[57], and comprehensive annotation files including structural annotations in GFF3 format and genomic sequences in FASTA format are provided via Figshare (https://doi.org/10.6084/m9.figshare.30164734)[58]. All raw sequencing data (HiFi, Hi-C, ONT, genome survey, and RNA-seq) generated in this study are available from NCBI Sequence Read Archive (SRA) under accession number SRP538810[59].

Code availability

No custom code was developed for this study. All genome assembly, annotation, and validation analyses were performed using publicly available bioinformatics software with standard protocols and default parameters, as described in the Methods section.

Received: 13 January 2025; Accepted: 24 November 2025;
Published online: 02 December 2025

References

1. Motomura, H. *Threadfins of the world (Family Polynemidae): An annotated and illustrated catalogue of polynemid species known to date*. (Food and Agricultural Organization, https://api.semanticscholar.org/CorpusID:128056442, 2004).
2. Ma, X. & Wang, W. X. Unveiling Osmoregulation and Immunological Adaptations in *Eleutheronema tetradactylum* Gills through High-Throughput Single-Cell Transcriptome Sequencing. *Fish & shellfish immunology.* **154**, 109878, https://doi.org/10.1016/j.fsi.2024.109878 (2024).
3. Motomura, H., Iwatsuki, Y., Kimura, S. & Yoshino, T. Revision of the Indo-West Pacific polynemid fish genus *Eleutheronema* (Teleostei: Perciformes). *Ichthyological Research.* **49**, 47–61, https://doi.org/10.1007/s102280200005 (2002).
4. Huang, C. T. *et al*. Bioeconomic evaluation of *Eleutheronema tetradactylum* farming: A case study in Taiwan. *Fisheries Science.* **88**, 437–447, https://doi.org/10.1007/s12562-022-01591-4 (2022).
5. IUCN. The IUCN Red List of Threatened Species. Version 2015-4. Available at: https://www.iucnredlist.org/species/46087646/57168342#bibliography (2015).
6. Moore, B. *et al*. Stock structure of blue threadfin *Eleutheronema tetradactylum* across northern Australia, as indicated by parasites. *Fish Biology.* **78**, 923–936, https://doi.org/10.1111/j.1095-8649.2011.02917.x (2011).
7. Xuan, Z. & Wang, W. X. Diversity of life history and population connectivity of threadfin fish *Eleutheronema tetradactylum* along the coastal waters of Southern China. *Scientific Reports.* **13**, 3976, https://doi.org/10.1038/s41598-023-31174-x (2023).
8. Azad, I. S., Al-Yaqout, A., El-Dakour, S., Kawahara, S. & Al-Roumi, M. First record of iridovirus (ISKNV) infections in Fourfinger threadfin from Kuwait. *Journal of King Saud University – Science.* **36**, 103393–103393, https://doi.org/10.1016/j.jksus.2024.103393 (2024).
9. Bharadhirajan, P. *et al*. Prevalence of copepod parasite (*Lernaeenicus polynemi*) infestation on *Eleutheronema tetradactylum* from Pazhayar coastal waters, southeast coast of India. *Journal of Coastal Life Medicine.* **1**, 258–261, https://doi.org/10.12980/JCLM.1.20133D154 (2013).
10. Abu Hena, M. K., Idris, M. H., Wong, S. K. & Kibria, M. M. Growth and survival of Indian salmon *Eleutheronema tetradactylum* (Shaw, 1804) in brackish water pond. *Journal of Fisheries and Aquatic Science.* **6**, 479–484, https://doi.org/10.3923/jfas.2011.479.484 (2011).
11. Iqbal, T. H. *et al*. Feeding habits of four-finger threadfin fish, *Eleutheronema tetradactylum*, and its diet interaction with co-existing fish species in the coastal waters of Thailand. *Peer J* **11**, e14688, https://doi.org/10.7717/peerj.14688 (2023).
12. Soe, K. K. *et al*. Reproductive characteristics of the hermaphroditic four-finger threadfin, *Eleutheronema tetradactylum* (Shaw, 1804), in tropical coastal waters. *BMC zoology* **8**, 22, https://doi.org/10.1186/s40850-023-00181-w (2023).
13. Jin, J. H., Amenyogbe, E., Yang, Y., Wang, Z. L. & Lu, Y. Effects of ammonia nitrogen stress on the physiological, biochemical, and metabolic levels of the gill tissue of juvenile four-finger threadfin (*Eleutheronema tetradactylum*). *Aquatic toxicology (Amsterdam, Netherlands).* **274**, 107049, https://doi.org/10.1016/j.aquatox.2024.107049 (2024).
14. Xiao, J., Tsim, K. W. K., Hajisamae, S. & Wang, W. X. Chromosome-level genome and population genomics provide novel insights into adaptive divergence in allopatric *Eleutheronema tetradactylum*. *International Journal of Biological Macromolecules.* **244**, 125299, https://doi.org/10.1016/j.ijbiomac.2023.125299 (2023).
15. Deng, Y. *et al*. A telomere-to-telomere gap-free reference genome of watermelon and its mutation library provide important resources for gene discovery and breeding. *Molecular Plant.* **15**, 1268–1284, https://doi.org/10.1016/j.molp.2022.06.010 (2022).
16. Zhang, Y. H. *et al*. The telomere-to-telomere genome of Fragaria vesca reveals the genomic evolution of Fragaria and the origin of cultivated octoploid strawberry. *Hortic Research.* **10**, uhad027, https://doi.org/10.1093/hr/uhad027 (2023).

17. Belton, J. M. *et al*. Hi–C: A comprehensive technique to capture the conformation of genomes. *Methods.* **58**, 268–276, https://doi.org/10.1016/j.ymeth.2012.05.001 (2012).
18. 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).
19. 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).
20. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications.* **11**, 1432, https://doi.org/10.1038/s41467-020-14998-3 (2020).
21. Cheng, H.-Y., 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).
22. Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC bioinformatics.* **19**, 1–10, https://doi.org/10.1186/s12859-018-2485-7 (2018).
23. Burton, J. N. *et al*. Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nature Biological.* **31**, 1119–1125, https://doi.org/10.1038/nbt.2727 (2013).
24. Kajitani, R. *et al*. Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research.* **24**, 1384–1395, https://doi.org/10.1101/gr.170720.113 (2014).
25. Durand, N. C. *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).
26. 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).
27. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* **34**, 3094–3100, https://doi.org/10.1093/bioinformatics/bty191 (2018).
28. Xu, M. Y. *et al*. TGS-GapCloser: A fast and accurate gap closer for large genomes with low coverage of error-prone long reads. *Gigascience.* **9**, giaa094, https://doi.org/10.1093/gigascience/giaa094 (2020).
29. 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).
30. Benson, G. J. N. 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).
31. Tarailo-Graovac, M. & Chen, N. S. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics*. **25**, 4. 10. 1-4. 10. 14, https://doi.org/10.1002/0471250953.bi0410s25 (2009).
32. Price, A. L., Jones, N. C. & Pevzner, P. A. *De novo* identification of repeat families in large genomes. *Bioinformatics.* **21**, i351–i358, https://doi.org/10.1093/bioinformatics/bti1018 (2005).
33. Ou, S. & Jiang, N. LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mobile DNA.* **10**, 48, https://doi.org/10.1186/s13100-019-0193-0 (2019).
34. Vij, S. *et al*. Chromosomal-Level Assembly of the Asian Seabass Genome Using Long Sequence Reads and Multi-layered Scaffolding. *PLOS Genetics.* **12**, e1005954, https://doi.org/10.1371/journal.pgen.1005954 (2016).
35. Shirak, A. *et al*. Identification of Repetitive Elements in the Genome of *Oreochromis niloticus*: Tilapia Repeat Masker. *Mar Biotechnol* **12**, 121–125, https://doi.org/10.1007/s10126-009-9236-8 (2010).
36. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics.* **6**, 1–11, https://doi.org/10.1186/1471-2105-6-31 (2005).
37. Shumate, A. & Salzberg, S. L. Liftoff: accurate mapping of gene annotations. *Bioinformatics.* **37**, 1639–1643, https://doi.org/10.1093/bioinformatics/btaa1016 (2021).
38. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research.* **34**, W435–W439, https://doi.org/10.1093/nar/gkl200 (2006).
39. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology.* **268**, 78–94, https://doi.org/10.1006/jmbi.1997.0951 (1997).
40. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology.* **33**, 290–295, https://doi.org/10.1038/nbt.3122 (2015).
41. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics.* **12**, 1–14, https://doi.org/10.1186/1471-2105-12-491 (2011).
42. Ashburner, M. *et al*. Gene ontology: tool for the unification of biology. *Nature Genetics.* **25**, 25–29, https://doi.org/10.1038/75556 (2000).
43. 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).
44. Boeckmann, B. *et al*. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research.* **31**, 365–370, https://doi.org/10.1093/nar/gkg095 (2003).
45. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research.* **28**, 45–48, https://doi.org/10.1093/nar/28.1.45 (2000).
46. Sayers, E. W. *et al*. Database resources of the national center for biotechnology information. *Nucleic Acids Research.* **50**, D20–D26, https://doi.org/10.1093/nar/gkg033 (2022).
47. Shen, W.-K. *et al*. AnimalTFDB 4.0: a comprehensive animal transcription factor database updated with variation and expression annotations. *Nucleic Acids Research.* **51**, D39–D45, https://doi.org/10.1093/nar/gkac907 (2023).
48. BuchfInk, B., Reuter, K. & Drost, H.-G. Sensitive protein a lignments at tree-of-life scale using DIAMOND. *Nature methods.* **18**, 366–368, https://doi.org/10.1038/s41592-021-01101-x (2021).
49. Bateman, A. *et al*. The Pfam Protein Families Database. *Nucleic Acids Research.* **28**, 263–266, https://doi.org/10.1093/nar/28.1.263 (2000).
50. Paysan-Lafosse, T. *et al*. InterPro in 2022. *Nucleic Acids Research.* **51**, D418–D427, https://doi.org/10.1093/nar/gkac993 (2023).
51. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics.* **30**, 1236–1240, https://doi.org/10.1093/bioinformatics/btu031 (2014).
52. Bu, D. H. *et al*. KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis. *Nucleic acids research.* **49**, W317–W325, https://doi.org/10.1093/nar/gkab447 (2021).
53. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology.* **215**, 403–410, https://doi.org/10.1016/S0022-2836(05)80360-2 (1990).
54. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research.* **25**, 955–964, https://doi.org/10.1093/nar/25.5.955 (1997).
55. Griffiths-Jones, S. *et al*. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Research.* **33**, D121–D124, https://doi.org/10.1093/nar/gki081 (2005).
56. Tang, H. *et al*. Synteny and Collinearity in Plant Genomes. *Springer Netherlands.* **320**, 486–488, https://doi.org/10.1126/science.1153917 (2008).
57. Zhang, H. J. *et al*. A chromosome-level genome assembly of the four-finger threadfin (*Eleut- heronema tetradactylum*). *GenBank* https://identifiers.org/ncbi/insdc:JBIEKL000000000 (2024).

58. Zhang, H. J., *et al*. A telomere-to-telomere genome of the fourfinger threadfin (*Eleutheronema tetradactylum*). *Figshare* (https://doi.org/10.6084/m9.figshare.30164734).
59. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP538810 (2024).
60. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Molecular biology and evolution*. **38**, 4647–4654, https://doi.org/10.1093/molbev/msab199 (2021).

Acknowledgements

The research was financially supported by the Guangdong Province Ordinary Colleges and Universities Key Field Special Project (Science and Technology Services for Rural Revitalization) (2023ZDZX4011), Guangdong Ocean University Aquaculture Excellent Young Talent Program (2024), and the Guangdong Province Ordinary Colleges and Universities Innovation Team Projects (2021KCXTD026; 2022KCXTD013).

Author contributions

These authors contributed equally: Huijuan Zhang, Yifei Pan, Benxun Miao, Linjuan Wang, Minxuan Jin, Anna Zheng, Jiandong Zhang, Baogui Tang, Jiansheng Huang, Jing Li, Dee-hwa Chong and Zhongliang Wang. H.J.Z. and L.J.W. conceived the project. Y.F.P., B.X.M., M.X.J., A.N.Z., J.D.Z., B.G.T., J.S.H., J.L., D.C. and Z.L.W. collected the samples. H.J.Z., Y.F.P., B.X.M., M.X.J., A.N.Z., J.D.Z., B.G.T., J.S.H., J.L., D.C. and Z.L.W. performed the genome assembly, gene annotation and other bioinformatics analysis. H.J.Z. and L.J.W. wrote and revised the manuscript. B.G.T., J.S.H., J.L. and Z.L.W. revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

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

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.