www.nature.com/scientificdata

scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly and annotation of Japanese anchovy (*Engraulis japonicus*)

Shufang Liu[1,2,6], Le Wang[3,6], Ruixiang Wang[1,4], Huan Wang[1], Ang Li[1,2], Changting An[1], Zining Meng[5] & Zhimeng Zhuang[1] ✉

The Japanese anchovy (*Engraulis japonicus*), a finfish with the largest biomass of a single species in the Yellow and East China Seas, plays an important pivotal role in converting zooplanktons into high trophic fish in the food web. As a result, the fish is regard as a key species in its habiting ecosystem. However, the lack of genomic resources hampers our understanding of its genetic diversity and differentiation, as well as the evolutionary dynamics. Here, we firstly report a complex chromosome-level genome assembly of *E. japonicus* with a large size of 1.4 Gb, with features of high repetitive sequences (54.9%), high heterozygosity (2.3%) and a number of protein-coding genes (24,405). The genome sequence exhibited a remarkable degree of completeness, valued 94.07% of the complete BUSCO. This work firstly reported the genome sequence of *E. japonicus*, offering the crucial resources for further studies on the genetic diversity and adaptive evolution of this species.

Background & Summary

Genomic resources, specifically genome sequences, are of particular importance in various genetic studies. Whole genome sequences are of help in examining the chromosomal evolution through comparative genomics, dissecting the genomic architecture for ecological adaptation, pinpointing the genes responsible for notable phenotypes as well as elucidating the divergence and speciation of organisms[1–3]. The technologies of high-throughput genome sequencing and cost-effective, precise genome assembly algorithms have promoted the assembly and release of numerous genome sequences, meanwhile, have substantially made the progress in genomics, offering comprehensive and novel insights into the fundamental mechanisms behind various biological questions of interest[4,5].

The Japanese anchovy (*Engraulis japonicus*) is a petite marine finfish belonging to the Clupeiformes order, distributing in the northwest Pacific marginal seas, northward from the Sea of Japan and southward to the East China Sea[6]. This anchovy with a great biomass in the region, plays a pivotal role in the food chain due to being as both a forage and a food fish[7]. During the late 1990s, its peak annual catch was about one million tons[8]. However, due to the high capture pressure and adverse effects of global climate change on marine ecosystem, its population size had substantially declining[8,9]. Unfortunately, the species has recently been classified as overexploited. Like some other migratory fish in the region such as *Larimichthys polyactis* and *L. crocea*[10], *E. japonicus* exhibits a migratory behaviour between spawning and overwintering grounds[11]. So far, the presence of genetic variation among different migratory stocks of *E. japonicus* remains controversial, primarily due to the use of different

[1]State Key Laboratory of Mariculture Biobreeding and Sustainable Goods, Yellow Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Qingdao, 266071, Shandong, China. [2]Laboratory for Marine Fisheries Science and Food Production Processes, Qingdao Marine Science and Technology Center, Qingdao, 266237, Shandong, China. [3]Molecular Population Genetics Group, Temasek Life Sciences Laboratory, 1 Research Link, National University of Singapore, Singapore, 117604, Singapore. [4]College of Fisheries and Life Science, Shanghai Ocean University, Shanghai, 201306, China. [5]State Key Laboratory of Biocontrol, Institute of Aquatic Economic Animals and the Guangdong Province Key Laboratory for Aquatic Economic Animals, School of Life Sciences, Sun Yat-sen University, Guangzhou, 510275, Guangdong, China. [6]These authors contributed equally: Shufang Liu, Le Wang. ✉e-mail: zhuangzm@ysfri.ac.cn

genetic markers and variations in the resolution of analytical methods[12–15]. Population genetic studies based on sequence variation in mitochondrial cytochrome b (*Cyt b*) and mitochondrial DNA control region fragments revealed no significant genetic structure across the wide-ranging populations of *E. japonicus* in the northwestern Pacific[12,13]. However, another molecular analyse using fragments of the *Cyt b* gene revealed considerable genetic variation among populations in the southern East China Sea[14]. Similarly, study utilizing six microsatellite loci detected weak but significant genetic differentiation between populations from the northeastern and southwestern coasts of Taiwan[15]. Marginally significant genetic differentiation was also observed between regional populations, such as the "Bohai Sea population (BHS)" and the "Japan Sea population (JPS)", as well as between the "North Yellow Sea population (NYS)" and the "Japan Sea population (JPS)" using restriction-site associated DNA sequencing (RADseq)[16]. As highlighted above, it should be noted that traditional approaches, which rely on limited genetic data from narrow genomic regions, may not fully capture the population structure of *E. japonicus*. The discrepancies between these studies may therefore hinder the accuracy and effectiveness of fisheries management and conservation efforts. Recently, genome scans based on the whole genome sequencing data have identified numerous loci under putative natural selection. These genetic loci, with significant genetic differentiation among stocks, can be utilized to assign the different stocks within a given population, which is helpful for management and conservation of fishery resources[10,16–18]. Understandably, these genomic resources are invaluable for those investigations like adaptive evolution, population dynamics, and genetic conservation etc.

Despite the ecological and commercial importance, the genomic features of this species remain unknown. The previous investigations were mostly concerted with the population structure identification by using microsatellite[15], and mitochondrial DNA markers[12,13], RADseq[16]. So far, there has not been existed any report about transcriptome or genome sequence datasets of this species. Moreover, genomic data for anchovy fish in general are limited, with genome sequences available for only six species, including *Coilia nasus*, *C. grayii*, *Encrasicholina punctifer*, *E. encrasicolus*, *Setipinna tenuifilis*, and *Thryssa baelama*. This scarcity has greatly hindered our understanding of the evolutionary processes and environmental adaptations within the Engraulidae family and even the broader Clupeiformes order.

To address this, we have utilized the Pacific Biosciences (PacBio) HiFi long-read, Hi-C (chromosome conformation capture), and Illumina short-read sequencing technologies to construct a high-quality chromosome-level genome sequence of the Japanese anchovy. Moreover, we conducted annotation and analysis of the genome in comparison with the related species. The workflow of *de novo* genome assembly and annotation is shown in the Fig. 1. The highly accurate, chromosome-level reference genome would promote the progress of both population genetics and evolutionary biology of this species, as well as make it possible for the comparative genomics studies among the species of Clupeiformes order.

Methods

Ethics statement.

All experiments were performed according to the Guidelines for the Care and Use of Laboratory Animals in China. All experimental procedures and sample collection methods were approved by the Institutional Animal Care and Use Committee (IACUC) of Yellow Sea Fisheries Research Institute, CAFS under approval No. YSFRI-2022041.

Sample collection and sequencing.

A mature female *E. japonicus* (Fig. 2) was obtained from the coastal waters of the Yellow Sea, close to Qingdao, China. Its dorsal muscle was collected for subsequently DNA extraction using a standard sodium dodecyl sulfate (SDS) extraction method. Subsequently, the concentration and quality of the extracted genomic DNA (gDNA) were quantified and assessed using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific) and by running a 0.8% agarose gel, respectively. The high-quality gDNA was initially employed to establish a short-insert library of approximately 350 bp using the TruSeq DNA PCR-Free kit (Illumina, USA). The library was subsequently sequenced on the Illumina NovaSeq 6000 platform (Illumina, USA), and approximately 101 Gb of 2 × 150 bp reads were generated (Table 1). Long-read sequencing was carried out on the same sample using the PacBio HiFi sequencing technology (Pacific Biosciences, USA). A standard PacBio library with an insert size of 20 kb was prepared using the SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences, USA). Subsequently, the library was sequenced on a PacBio Sequel II system (Pacific Biosciences, USA), yielding a total of 51.3 Gb of PacBio HiFi reads, with an N50 length of 17.4 kb (Table 1). Lastly, a Hi-C library was established according to a previous protocol[19] with some modifications[20]. In summary, muscle samples from the same sequenced individual were cross-linked using 4% formaldehyde. The fixed samples were then homogenized to isolate the nuclei. Following that, the DNA was digested with the MboI restriction enzyme (NEB, USA). The digested products underwent sequential treatments for end repairing, biotin labelling, and ligation of blunt-end fragments. The ligated DNA was subsequently sheared into fragments with a peak size of 400 bp. These fragments were then used to construct a standard DNA library using the TruSeq DNA Sample Prep Kit (Illumina, USA). The Hi-C library was sequenced for 2 × 150 bp reads on the Illumina NovaSeq 6000 platform, generating a total of 109.5 Gb reads (Table 1).

For transcriptome sequencing, samples of the brain, ovary, heart, muscle, and liver were obtained from the same sequenced sample for RNA extraction, using TRIzol™ Reagent (Thermo Fisher Scientific, USA). The concentration and quality of the total RNA were quantified and evaluated utilizing a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, USA) and by running a 1.0% agarose gel, respectively. Total RNA from each individual sample was employed to construct mRNA libraries using the TruSeq RNA Library Prep Kit v2 (Illumina, USA). Subsequently, the libraries were sequenced on the Illumina NovaSeq 6000 platform (Illumina, USA), yielding an average of 5.58 Gb of 2 × 150 bp reads for each transcriptome sample (Table 1).

Chromosome-level genome assembly.

The Illumina reads were first cleaned using the program NGSQCToolkit v2.3[21]. The cleaned reads were then utilized to estimate genome parameters based on the 17-mer

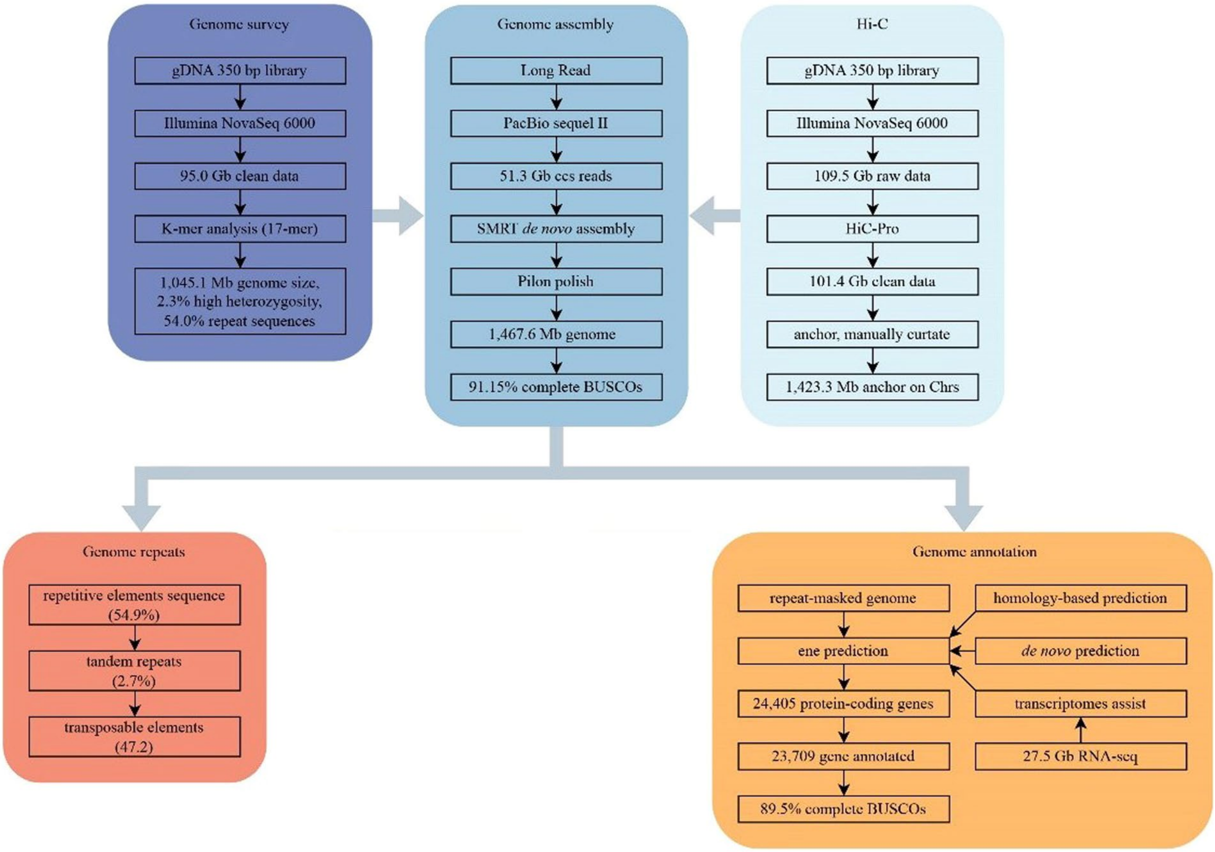

Fig. 1 The overview of the chromosome-level genome assembly and annotation. Chrs: chromosomes. We first used 95.0 Gb short-read sequencing data to predict the assembled genome size was approximately 1,045.1 Mb by K-mer analysis, and the repeat sequences and heterozygosity were approximately 54.0% and 2.3%, respectively. Then, the 51.3 Gb of PacBio ccs data resulted in a 1,467.6 Mb assembly, with contig N50 of 456.3 kb. The contigs were anchored into 24 pseudo-chromosomes covering roughly 95.2% of the genome assembly with the assistance of 109.5 Gb Hi-C reads. The final assembly consisted of 24 pseudo-chromosomes that yielded 1,423.3 Mb of *E. japonicus* genome, with a scaffold N50 of 55.0 Mb. The genome contained 54.9% repeat sequences and 23,709 genes were functionally annotated from a total of 24,405 (97.15%) predicted protein-coding genes by combination of RNAseq and ISO-Seq annotation, genome sequence, and homolog protein.

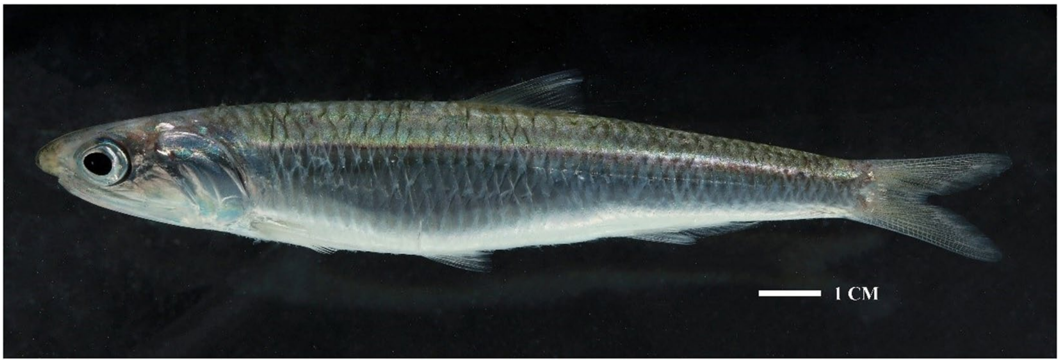

Fig. 2 The mature female Japanese anchovy (*Engraulis japonicus*) obtained from the coastal waters of the Yellow Sea.

frequency distribution using the program GenomeScope v2.09[22]. The estimated genome size, heterozygosity, and content of repetitive sequences were found to be 1,045.1 Mb, 2.3%, and 54.0%, respectively. Subsequently, the Pacbio HiFi reads were assembled into contigs using the program Hifiasm v0.19.5[23], with default parameters. The assembled contigs were then polished using Pilon v1.22[24], also with default parameters. The total length and N50 of the assembled contigs were approximately 1,467.6 Mb and 456.3 kb, respectively (Table 2).

To achieve a chromosome-level assembly, raw Hi-C sequencing reads were first filtered using HiC-Pro v2.8.0[25]. Subsequently, the cleaned reads were employed to anchor the assembled contigs into scaffolds using

Libraries	Insert size (bp)	Raw data (Gb)	Clean data (Gb)	Read length (bp)
Illumina reads	350	101.0	95.0	2 × 150
PacBio Hifi reads	20,000	51.3	50.3	17,435
Hi-C reads	350	109.5	101.4	2 × 150
Brain RNAseq	220	6.2	5.9	2 × 150
Ovary RNAseq	220	6.2	5.8	2 × 150
Heart RNAseq	220	6.2	5.9	2 × 150
Muscle RNAseq	220	5.0	4.8	2 × 150
Liver RNAseq	220	5.4	5.1	2 × 150

Table 1. Summary statistics of sequencing libraries and reads used in this study.

Features	Contigs	Scaffolds
Total number	6,342	386
Total length (bp)	1,467,598,116	1,423,324,225
Max length (bp)	3,423,671	69,055,323
N50 length (bp)	456,289	55,013,877
N50 number	872	13
N90 length (bp)	94,195	49,130,702
N90 number	3,553	24

Table 2. Summary statistics of the assembled contigs and scaffolds of *Engraulis japonicus*.

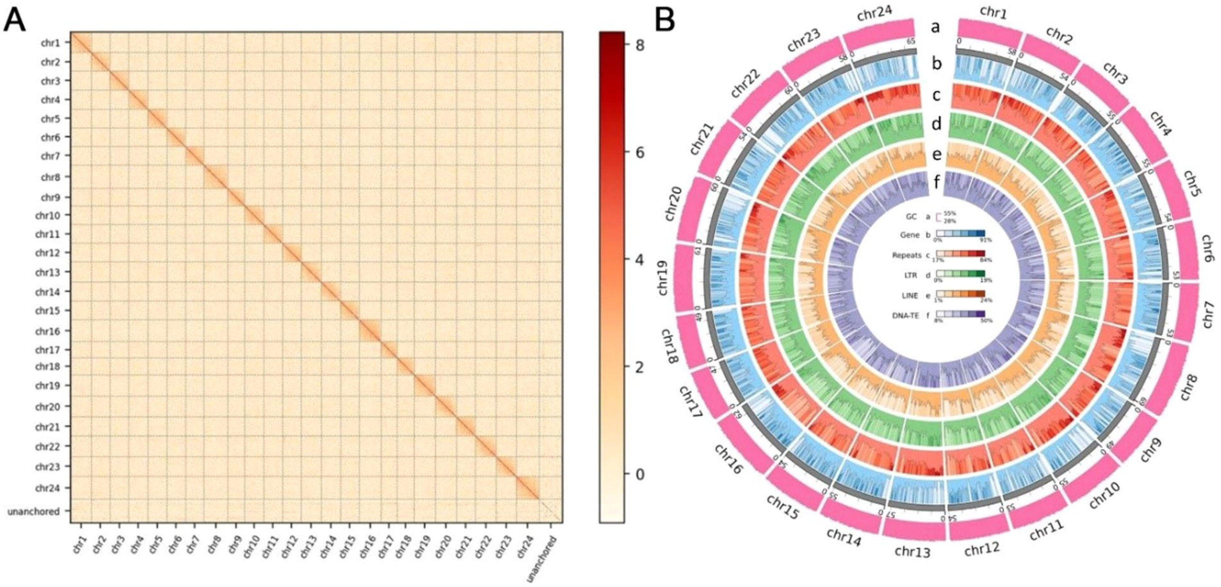

Fig. 3 Chromosome-level assembly and features of the *Engraulis japonicus* genome. (**A**) Genome-wide chromatin interactions in the *E. japonicus* genome revealed by heatmap. (**B**) Circos plot of genomic features in the *E. japonicus* genome in a 100-kb window size. Each circle from outside to inside represents GC content along individual pseudochromosomes with indicated length (a), gene density (b), density of repetitive sequences (c), density of LTR elements (d), density of LINE elements (e) and density of DNA transposable elements (f).

Juicer[26] and 3D-DNA pipelines[19]. The assembled scaffolds were then manually curated using Juicebox[27], with a prior setting of 24 haploid chromosomes[28]. Consequently, 95.2% of the assembled contigs were anchored to 24 pseudochromosomes (Fig. 3A), with individual chromosome lengths ranging from 47.0 Mb to 69.1 Mb (Fig. 3B and Table 3). The total length of the chromosome-level genome assembly amounted to 1,423.3 Mb, with a scaffold N50 of 55.0 Mb (Table 2). This discrepancy in genome assembly size, as opposed to the previously mentioned prediction, can be attributed to the tendency of short-read sequencing to underestimate the size of highly repetitive and heterozygous genomes[29].

Repetitive sequence annotation. Annotations of repetitive sequences were conducted using Repeatmasker v4.0.6[30], based on the RepBase database v202101[31] and a custom repeat library. The custom repeat

Chr.	Length (bp)	Percentage (%)
chr1	58,809,496	4.1
chr2	54,133,493	3.8
chr3	54,955,785	3.9
chr4	55,034,740	3.9
chr5	54,604,410	3.8
chr6	53,828,334	3.8
chr7	53,569,943	3.8
chr8	69,055,323	4.9
chr9	49,547,565	3.5
chr10	55,013,877	3.9
chr11	53,544,828	3.8
chr12	54,548,640	3.8
chr13	57,688,500	4.1
chr14	55,739,507	3.9
chr15	54,528,944	3.8
chr16	62,822,442	4.4
chr17	47,030,922	3.3
chr18	49,130,702	3.5
chr19	61,740,138	4.3
chr20	60,672,501	4.3
chr21	54,602,669	3.8
chr22	60,061,282	4.2
chr23	58,316,196	4.1
chr24	65,693,204	4.6
unchr	68,650,784	4.8

Table 3. Summary statistics of the length of pseudochromosomes of *Engraulis japonicus*.

	RepBase TEs	TE Proteins	De novo	Combined TEs	
Classifications	Length (bp)	Length (bp)	Length (bp)	Length (bp)	Percentage (%)
DNA	281,372,463	12,262,076	205,887,363	439727956	30.9
LINE	54,790,448	24,396,415	46,786,211	89305039	6.3
SINE	6,479,834	0	7,481,582	12418808	0.9
LTR	71,416,312	15,715,349	62,531,367	129152198	9.1
Satellite	30,019,921	0	8,537,638	37951176	2.7
Simple_repeat	0	0	31,947	31947	0
Other	20,283	0	0	20283	0
Unknown	6,122,332	9,318	170,712,575	176666970	12.4
Total	372,141,572	52,358,445	488,125,606	780865692	54.9

Table 4. Summary statistics of the predicted sequence repeats in the assembled genome of *Engraulis japonicus*.

library was generated utilizing RepeatModeler v2.0.5[32], with default parameters. Additionally, the programs LTR_FINDER v1.06[33] and Tandem Repeat Finder v4.07[34] were independently employed to identify long terminal repeats and tandem repeats, using default parameters. The predictions of these programs were then consolidated to create a nonredundant library of repetitive sequences within the genome, which was subsequently used for annotation within Repeatmasker. A total of 780.9 Mb, constituting 54.9% of the assembled genome, were annotated as repetitive sequences (Table 4). Among these repeats, long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), and long terminal repeats (LTRs) accounted for 6.3%, 0.9%, and 9.1% of the genome, respectively (Table 4).

Gene prediction and functional annotation. Predictions of protein-coding genes were carried out on a repeat-masked genome utilizing homology-, evidence- and ab initio-based prediction methods. For the homology-based gene prediction, protein sequences of *Alosa alosa* (GCF_017589495.1), *A. sapidissima* (GCF_018492685.1), *S. tenuifilis* (v1)[35], *C. nasus* (v1)[36], and *Danio rerio* (NCBI, GCF_000002035.6) were aligned to the *E. japonicus* genome assembly using BLASTP v2.2.24[37] with default parameters. Regarding evidence-based annotation, the mentioned transcriptomes were assembled utilizing Trinity v2.1.1[38] with default parameters, and then condensed into a nonredundant transcript dataset for utilization as supporting evidence for prediction. The Maker v2.53 pipeline[39] was employed to consolidate the predictions from both

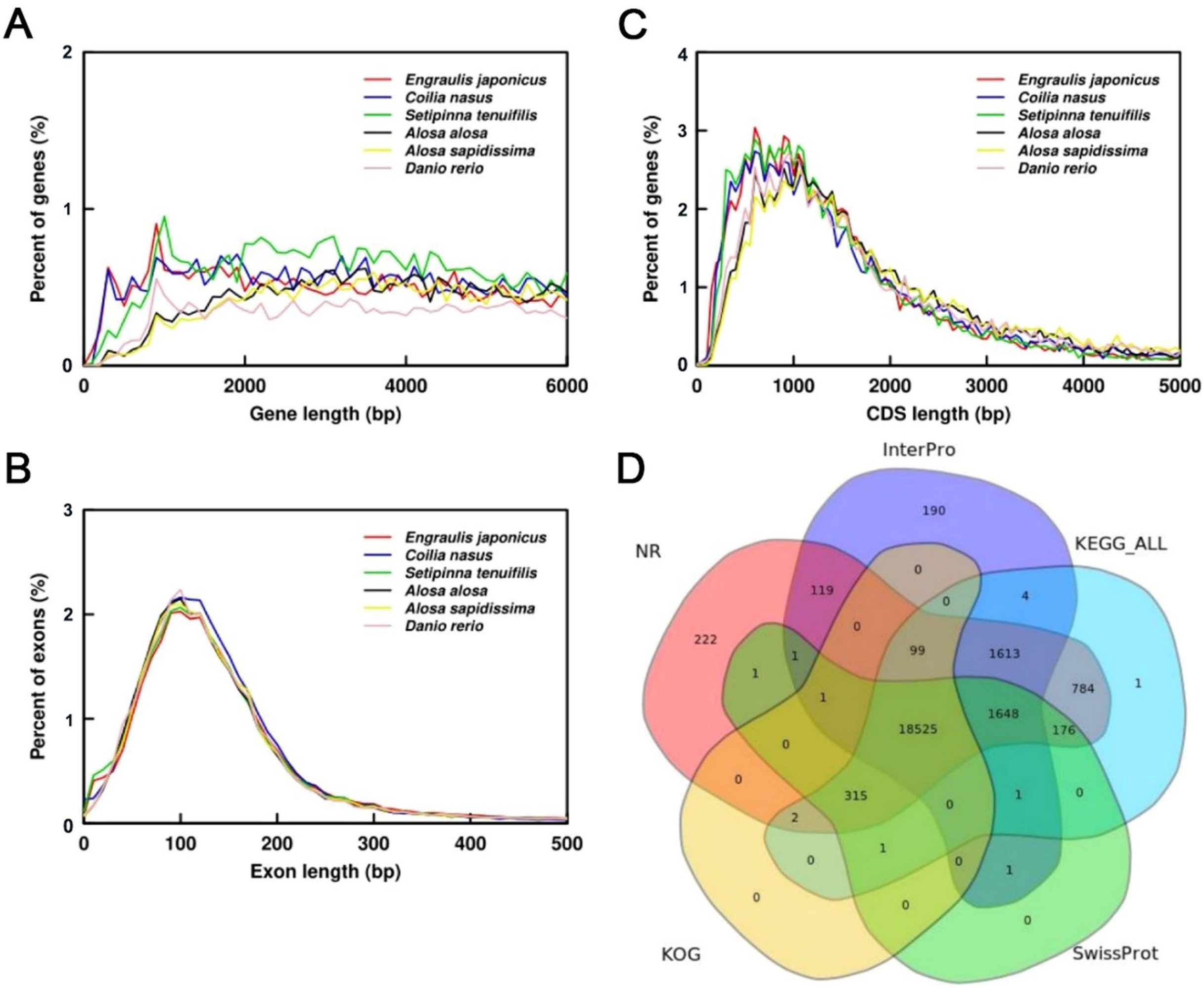

Fig. 4 Features of the predicted protein coding genes in the *Engraulis japonicus* genome. (**A**) Distribution of the length of genes among six studied species. (**B**) Distribution of the length of exons. (**C**) Distribution of the length of coding sequences (CDS). (**D**) Summary statistics of the number of genes annotated by different databases: NR, InterPro, KEGG, KOG and SwissProt.

	Gene		mRNA	
Gene/Database	Number	Percent (%)	Number	Percent (%)
Total	24,405	100	32,891	100
Annotated	23,709	97.15	31,830	96.77
NR	23,506	96.32	31,571	95.99
SwissProt	20,670	84.7	28,058	85.31
TrEMBL	23,407	95.91	31,450	95.62
KOG	18,943	77.62	25,849	78.59
TF	5,432	22.26	7,331	22.29
InterPro	22,202	90.97	29,764	90.49
GO	16,628	68.13	22,251	67.65
KEGG_ALL	23,169	94.94	31,117	94.61
KEGG_KO	16,500	67.61	22,413	68.14
Pfam	20,745	85	27,776	84.45
Unannotated	696	2.85	1,061	3.23

Table 5. Summary statistics of the numbers of predicted protein coding genes in the assembled genome of *Engraulis japonicus*.

the homology- and evidence-based approaches. Predicted gene models were iteratively trained using SNAP v2006.07.28[40], GeneMark-EP v4.72[41], and Augustus v3.3.2[42] for three iterations. Subsequently, predicted gene

Type	Copy number	Average length (bp)	Total length (bp)	Percentage of genome (%)
miRNA	1,492	89	132,236	0.009291
tRNA	19,120	75	1,442,638	0.101357
rRNA	229	175	40,174	0.002823
snRNA	3,143	152	477,202	0.033527

Table 6. Summary statistics of noncoding RNAs in the genome assembly of *Engraulis japonicus*.

Type	Proteins	Percentage (%)
Complete BUSCOs (C)	3,424	94.07
Complete and single-copy BUSCOs (S)	3,229	88.71
Complete and duplicated BUSCOs (D)	195	5.36
Fragmented BUSCOs (F)	62	1.7
Missing BUSCOs (M)	153	4.2
Total BUSCO groups searched	3,640	100

Table 7. Assessment of the completeness of the genome assembly of *Engraulis japonicus* using BUSCO.

Clupeiform species	Accession	Genome size	Contigs N50	Scaffolds N50	Assembly level	Complete BUSCO
Alosa alosa	GCA_017589495.2	854.4 Mb	1.2 Mb	35.4 Mb	Chromosome	91.5%
Alosa fallax	GCA_029875135.1	755.9 Mb	2.0 kb	2.1 kb	Scaffold	—
Alosa sapidissima	GCA_018492685.1	903.6 Mb	1.6 Mb	38.4 Mb	Chromosome	95.6%
Clupea harengus	GCA_900700415.2	786.3 Mb	1.0 Mb	29.8 Mb	Chromosome	94.9%
Coilia grayii	GCA_042479465.1	920.6 Mb	2.6 Mb	36.4 Mb	Chromosome	—
Coilia nasus	GCA_027475355.1	851.7 Mb	26.6 Mb	35.4 Mb	Chromosome	92.5%
Denticeps clupeoides	GCA_900700375.2	567.4 Mb	3.1 Mb	22.8 Mb	Chromosome	94.6%
Encrasicholina punctifer	GCA_041295995.1	1.2 Gb	2.2 kb	21.9 Mb	Scaffold	—
Engraulis encrasicolus	GCA_034702125.1	1.4 Gb	1.4 Mb	56.4 Mb	Chromosome	90.4%
Limnothrissa miodon	GCA_017657215.1	580.9 Mb	31.1 kb	455.5 kb	Scaffold	—
Nematalosa erebi	GCA_036025975.1	757.2 Mb	2.9 kb	—	Contig	—
Sardina pilchardus	GCA_963854185.1	869.4 Mb	1.1 Mb	34.6 Mb	Chromosome	84.5%
Sardinella fimbriata	GCA_030264415.1	899.5 Mb	1.7 kb	4.1 kb	Scaffold	—
Sardinella gibbosa	GCA_034783695.1	812.5 Mb	2 kb	7.1 kb	Scaffold	—
Sardinella hualiensis	GCA_032353275.1	709.4 Mb	2.6 kb	11.4 Mb	Scaffold	—
Sardinella lemuru	GCA_030264355.1	545.3 Mb	1.4 kb	31.5 kb	Scaffold	—
Sardinella longiceps	GCA_027497455.2	1.1 Gb	153.4 kb	31.1 Mb	Scaffold	—
Sardinella tawilis	GCA_030264315.1	783.3 Mb	5 kb	14.9 Mb	Scaffold	—
Setipinna tenuifilis	GCA_030347295.1	798.4 Mb	1.4 Mb	32.4 Mb	Chromosome	89.6%
Sprattus sprattus	GCA_963457725.1	840.3 Mb	1.2 Mb	33.8 Mb	Chromosome	95.7%
Tenualosa ilisha	GCA_015244755.2	146.3 Mb	43.4 kb	—	Contig	—
Tenualosa thibaudeaui	GCA_027481765.1	668.9 Mb	17.1 Mb	—	Contig	—
Thryssa baelama	GCA_041045025.1	884.4 Mb	2.3 kb	5.7 kb	Scaffold	—
Engraulis japonicus	GCA_040112795.1	1.4 Gb	456 kb	55.0 Mb	Chromosome	94.07%

Table 8. Comparison of the genome assemblies of Clupeiform species.

models containing transposable element (TE) domains and lacking support from transcripts were filtered out and removed. As a result, a total of 24,405 nonredundant protein-coding genes were predicted. Upon comparing the gene set of *E. japonicus* with that of *A. alosa*, *A. sapidissima*, *S. tenuifilis*, *C. nasus*, and *D. rerio*, a similar distribution pattern in the length of genes (Fig. 4A), exons (Fig. 4B), and coding sequences (CDS) (Fig. 4C) was observed among these studied fish species.

Additionally, all predicted genes were functional annotated by mapping to the public databases including SwissProt, Nr, KEGG, and InterPro, COG, KOG, and Pfam. In total, 23,709 genes were classified by at least one of these databases, accounting for 97.1% of all the predicted protein coding genes in the *E. japonicus* genome (Table 5 and Fig. 4D). Furthermore, genes coding for tRNA were predicted using tRNAscan-SE v1.3.1[43] with default parameters. Genes for rRNA were predicted by aligning to invertebrate template rRNA sequences using BLASTN v2.2.24[37] with an E-value of 1e-5. Genes for both snRNAs and miRNAs were then identified using INFERNAL v1.1.1[44] against the Rfam database (release 12.0). In total, 23,984 non-coding RNAs (ncRNAs) were predicted, including 19,120 tRNAs, 229 rRNAs, 1,492 miRNAs, and 3,143 snRNAs (Table 6).

Data	Items	Result
PacBio HiFi long reads	Reads mapping rate (%)	99.91
	Genome average sequencing depth (×)	31.92
	Coverage of genome (%)	99.95
	Coverage of genome >4 × (%)	95.48
	Coverage of genome >10 × (%)	81.79
	Coverage of genome >20 × (%)	61.23
Illumina short reads	Reads mapping rate (%)	97.97
	Genome average sequencing depth (×)	64.36
	Coverage of genome (%)	99.60
	Coverage of genome > 4 × (%)	97.79
	Coverage of genome > 10 × (%)	92.39
	Coverage of genome > 20 × (%)	84.66

Table 9. Coverage statistics of PacBio HiFi long reads and Illumina short reads.

Data Records

All raw sequencing data are available on the NCBI through Bioproject PRJNA1082877[45]. The genome assembly and annotations are available on figshare[46] and the CNGB with accession number CNP0005377[47]. The assembled genome is also available on NCBI GenBank under the accession number GCA_040112795.1[48].

Technical Validation

Evaluation of the genome assembly.

To evaluate the quality of the genome assembly, the completeness of the genome sequence was first assessed by mapping to the Actinopterygii database (actinopterygii_odb10) of Benchmarking Universal Single-Copy Orthologs (BUSCO, v5.7.1). The genome assembly exhibited a high level of completeness, with a complete BUSCO value of 94.07%. Within this value, 88.71% were complete and single-copy while 5.36% were complete and duplicated. Only 1.7% BUSCOs were fragmented, and 4.2% were missing from the genome assembly (Table 7). We retrieved the genome assemblies of Clupeiformes archived in NCBI and found only 23 species with available genome sequences, of which only 10 species had chromosome-level genome assemblies (Table 8). The complete BUSCO value of *E. japonicus* (94.07%) is comparable to that of the high-quality chromosome-level genome assemblies of Clupeiform species archived in NCBI, which range from 84.5% to 95.6% with a median value of 92% (Table 8). Furthermore, both the PacBio HiFi long reads and Illumina short reads were aligned to the genome assembly using minimap2. The mapping rates for PacBio and Illumina reads were 99.91% and 97.97%, respectively (Table 9). Finally, the consensus quality value (QV), representing per-base consensus accuracy, was estimated using Merqury (v1.3), resulting in a QV of 49.74. Considering these data collectively, it is evident that the genome assembly of *E. japonicus* is characterized by both high completeness and high quality.

Code availability

No custom codes or scripts were utilized in this study. All bioinformatics programs and pipelines were executed according to the instructions and guidelines provided by the software developers. The specific software versions and corresponding parameters employed have been delineated in the Methods subsection.

Received: 31 May 2024; Accepted: 6 January 2025;
Published online: 22 January 2025

References

1. Wang, L. *et al.* A chromosome-level reference genome of african oil palm provides insights into its divergence and stress adaptation. *Genomics, Proteomics & Bioinformatics* **21**, 440–454 (2023).
2. Wang, L. *et al.* Genomic basis of striking fin shapes and colors in the fighting fish. *Molecular Biology and Evolution* **38**, 3383–3396 (2021).
3. Yue, G. & Wang, L. Current status of genome sequencing and its applications in aquaculture. *Aquaculture* **468**, 337–347 (2017).
4. Phillippy, A. M. New advances in sequence assembly. *Genome Research* **27**, xi–xiii (2017).
5. Jackson, S. A., Iwata, A., Lee, S. H., Schmutz, J. & Shoemaker, R. Sequencing crop genomes: approaches and applications. *New Phytologist* **191**, 915–925 (2011).
6. Takasuka, A. & Aoki, I. Environmental determinants of growth rates for larval Japanese anchovy *Engraulis japonicus* in different waters. *Fisheries Oceanography* **15**, 139–149 (2006).
7. Iversen, S., Zhu, D., Johannessen, A. & Toresen, R. Stock size, distribution and biology of anchovy in the Yellow Sea and East China Sea. *Fisheries Research* **16**, 147–163 (1993).
8. Yu, H. *et al.* Potential environmental drivers of Japanese anchovy (*Engraulis japonicus*) recruitment in the Yellow Sea. *Journal of Marine Systems* **212**, 103431 (2020).
9. Nakayama, S. I., Takasuka, A., Ichinokawa, M. & Okamura, H. Climate change and interspecific interactions drive species alternations between anchovy and sardine in the western North Pacific: Detection of causality by convergent cross mapping. *Fisheries Oceanography* **27**, 312–322 (2018).
10. Wang, L., Liu, S., Yang, Y., Meng, Z. & Zhuang, Z. Linked selection, differential introgression and recombination rate variation promote heterogeneous divergence in a pair of yellow croakers. *Molecular Ecology* **31**, 5729–5744 (2022).
11. Tanaka, H., Ohshimo, S., Takagi, N. & Ichimaru, T. Investigation of the geographical origin and migration of anchovy *Engraulis japonicus* in Tachibana Bay, Japan: A stable isotope approach. *Fisheries Research* **102**, 217–220 (2010).

12. Liu, J. X. *et al.* Late Pleistocene divergence and subsequent population expansion of two closely related fish species, Japanese anchovy (*Engraulis japonicus*) and Australian anchovy (*Engraulis australis*). *Molecular Phylogenetics and Evolution* **40**, 712–723 (2006).
13. Zheng, W., Zou, L. & Han, Z. Genetic analysis of the populations of Japanese anchovy *Engraulis japonicus* from the Yellow Sea and East China Sea based on mitochondrial cytochrome b sequence. *Biochemical Systematics and Ecology* **58**, 169–177 (2015).
14. Chen, C. S., Tzeng, C. H. & Chiu, T. S. Morphological and molecular analyses reveal separations among spatiotemporal populations of anchovy (*Engraulis japonicus*) in the southern East China Sea. *Zoological Studies* **49**, 270–282 (2010).
15. Yu, H. T., Lee, Y. J., Huang, S. W. & Chiu, T. S. Genetic analysis of the populations of Japanese anchovy (Engraulidae: *Engraulis japonicus*) using microsatellite DNA. *Marine Biotechnology* **4**, 471–479 (2002).
16. Zhang, B. D., Li, Y. L., Xue, D. X. & Liu, J. X. Population genomics reveals shallow genetic structure in a connected and ecologically important fish from the Northwestern Pacific Ocean. *Frontiers in Marine Science* **7**, 374 (2020).
17. Wang, L. *et al.* Population genetic studies revealed local adaptation in a high gene-flow marine fish, the small yellow croaker (*Larimichthys polyactis*). *PLoS One* **8**, e83493 (2013a).
18. Wang, L., Liu, S., Zhuang, Z., Lin, H. & Meng, Z. Mixed-stock analysis of small yellow croaker *Larimichthys polyactis* providing implications for stock conservation and management. *Fisheries Research* **161**, 86–92 (2015).
19. Dudchenko, O. *et al. De novo* assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
20. Wang, L. *et al.* A chromosome-level genome assembly of chia provides insights into high omega-3 content and coat color variation of its seeds. *Plant Communications* **3**, 100326 (2022a).
21. Patel, R. K. & Jain, M. NGSQCToolkit: a toolkit for quality control of next generation sequencing data. *PLoS One* **7**, e30619 (2012).
22. 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 (2020).
23. 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 (2021).
24. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).
25. Servant, N. *et al.* HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biology* **16**, 259 (2015).
26. Durand, N. C. *et al.* Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Systems* **3**, 95–98 (2016).
27. Durand, N. C. *et al.* Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Systems* **3**, 99–101 (2016).
28. Jinxing, W., Xiaofan, Z., Xiangmin, W. & Mingcheng, T. Karyotype analysis for seven species of clupeiform and perciform fishes. *Zoological Research* **15**, 76–79 (1994).
29. Pflug, J. M., Holmes, V. R., Burrus, C., Johnston, J. S. & Maddison, D. R. Measuring genome sizes using read-depth, k-mers, and flow cytometry: methodological comparisons in beetles (Coleoptera). *G3: Genes, Genomes, Genetics* **10**, 3047–3060 (2020).
30. Chen, N. Using Repeat Masker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics* **5**, 4.10. 11–14.10. 14 (2004).
31. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research* **110**, 462–467 (2005).
32. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences* **117**, 9451–9457 (2020).
33. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research* **35**, W265–W268 (2007).
34. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**, 573–580 (1999).
35. Liu, B. *et al.* Chromosome-level genome assembly and population genomic analysis reveal evolution and local adaptation in common hairfin anchovy (*Setipinna tenuifilis*). *Molecular Ecology* **00**, 1–18 (2023).
36. Xu, G. *et al.* Genome and population sequencing of a chromosome-level genome assembly of the Chinese tapertail anchovy (*Coilia nasus*) provides novel insights into migratory adaptation. *GigaScience* **9**, giz157 (2020).
37. McGinnis, S. & Madden, T. L. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research* **32**, W20–W25 (2004).
38. Grabherr, M. G. *et al.* Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature Biotechnology* **29**, 644–652 (2011).
39. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
40. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
41. Brůna, T., Lomsadze, A. & Borodovsky, M. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genomics and Bioinformatics* **2**, lqaa026 (2020).
42. Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research* **34**, W435–W439 (2006).
43. Chan, P. P. & Lowe, T. M. tRNAscan-SE: searching for tRNA genes in genomic sequences. *Methods in Molecular Biology* **1962**, 1–14 (2019).
44. Nawrocki, E. P., Kolbe, D. L. & Eddy, S. R. Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25**, 1335–1337 (2009).
45. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP492930 (2024).
46. Liu, S. *et al.* Chromosome-level genome assembly and annotation of Japanese anchovy (Engraulis japonicus). *figshare* https://doi.org/10.6084/m9.figshare.25273354 (2024).
47. *CNGB* https://db.cngb.org/search/project/CNP0005377/ (2024).
48. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_040112795.1 (2024).

Acknowledgements

This research was supported by the Project of Laoshan Laboratory (LSK202203802); the National Natural Science Foundation of China (Grant No. 42076132 and 32102768); and the China Agriculture Research System of MOF and MARA (CARS-47).

Author contributions

S.L. and L.W. conceived and designed this study and drafted the manuscript. S.L., Z.M. and Z.Z. coordinated and supervised the whole study. L.W., R.W. and H.W. conducted the genome assembly and bioinformatics analysis. R.W. and H.W. participated in manuscript improvement. A.L. and C.A. prepared the samples and the figures. S.L. and Z.Z. reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

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

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

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