scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly with telomeric repeats at scaffold ends for *Rhabdosargus sarba*

Mudagandur S. Shekhar, Vinaya Kumar Katneni ✉, Ashok Kumar Jangam, Karthic Krishnan, Sudheesh K. Prabhudas, Jesudhas Jani Angel Raymond & Kuldeep K. Lal

***Rhabdosargus sarba*, the goldlined seabream, is a euryhaline marine fish of great aquaculture potential. Genome sequencing and assembly of *R. sarba* was carried utilizing a multi-platform sequencing strategy that included long-read sequencing (PacBio HiFi), short-read sequencing (Illumina), and chromatin interaction mapping (Hi-C). The final genome assembly size after scaffolding was 764.59 Mb in 31 scaffolds with an N50 length of 33.98 Mb. Repeat profiling of primary assembly showed that 28.71% of the genome comprises of repeat elements. Gene prediction utilising the evidence from *ab initio* prediction and transcriptome data revealed 26,913 protein encoding genes and functional annotation and pathway analysis showed their participation in 332 pathways. This genome is an excellent resource for future research on genetic improvement and molecular breeding programmes for *R. sarba*.**

Background & Summary

Rhabdosargus sarba, also known as the goldlined seabream, is a species of marine ray-finned fish which belongs to the family Sparidae (Fig. 1). It is one of the commonly preferred food fishes with high domestic market demand in India. Being a euryhaline species, *R. sarba* has a potential for farming both in the ponds and cages. The successful broodstock development and breeding of *R. sarba* in captivity in India opens a new vista and adds a new species for research with an aim to harness commercial aquaculture potential. At molecular level, currently, research on goldlined seabream mainly is limited to biology and genetic studies. For example, SNP data was generated to explore the genetic adaptive evolution mechanisms[1], molecular phylogeny studies using mitochondrial genes[2], and mitogenome[3]. The limited reports on molecular studies, underscores the necessity for deciphering whole genome sequence of this fish that would encapsulate a broader spectrum of genetic information of this species at the genome level.

Currently there are twelve genomes available at NCBI's Genome database for the family Sparidae with genome assembly at chromosome and complete levels, excluding the genomes at contigs and scaffold levels. The genomes of *Acanthopagrus schlegelii* and *Pagrus major* assembled in 24 sequences with scaffold N50 of 31.8 Mb and 33.7 Mb respectively (Table 1). In the present study, Pacbio HiFi long reads along with HiC reads were combined to generate a high-quality genome of *R. sarba* assembled into 31 scaffolds with a total length of 764,592,746 bp and scaffold N50 of 33.98 Mb. The assembly is found to be at chromosome level with 24 longest scaffolds corresponding to chromosomes have telomeric ends and account for a total size of 764,366,029 bp. The current genome of *R. sarba* along with the genome of *E. suratensis*[4] could be assembled at chromosome scale with telomeres as compared to Pacbio Continuous Long Read (CLR) based assemblies for Asian seabass[5], Grey mullet[6] and Red snapper[7] and single-tube Long Fragment Read based assembly for grey mullet[8] indicating the superiority of Pacbio HiFi reads in assembling high quality chromosome level genomes for commercially important brackishwater fishes. The availability of this high-quality reference level genome of *Rhabdosargus sarba* will be a significant achievement in advancing our understanding of this fish species. It helps in paving the way for future studies on biological chracteristics, selective breeding programmes, broodstock management, ecological interactions, adaptive evolution and phylogenetic relationships of *R. sarba*.

ICAR-Central Institute of Brackishwater Aquaculture, No 75, Santhome High Road, MRC Nagar, Chennai, 600028, Tamil Nadu, India. ✉e-mail: vinayndri@yahoo.com

Fig. 1 A specimen of goldlined seabream (*Rhabdosargus sarba*).

S. no.	Assembly Accession	Assembly Name	Organism Name	Total Sequence Length	Contig N50	Scaffold N50	Number of Scaffolds
1	GCA_041753875.1	ASM4175387v1	*Acanthopagrus schlegelii*	714,975,658	31,795,397	31,795,397	24
2	GCA_040436345.1	Pma_NU_1.0	*Pagrus major*	785,972,034	8,664,321	33,667,259	24
3	GCA_900880675.2	fSpaAur1.2	*Sparus aurata*	833,578,411	2,862,625	35,791,275	175
4	GCA_900880675.1	fSpaAur1.1	*Sparus aurata*	833,578,411	2,862,625	35,791,275	175
5	GCF_900880675.1	fSpaAur1.1	*Sparus aurata*	833,578,411	2,862,625	35,791,275	175
6	GCA_904848185.1	fAcaLat1.1	*Acanthopagrus latus*	685,127,588	14,880,455	30,717,905	65
7	GCF_904848185.1	fAcaLat1.1	*Acanthopagrus latus*	685,127,588	14,880,455	30,717,905	65
8	GCA_963678695.2	Dpuntazzo_v2	*Diplodus puntazzo*	788,392,964	5,518,951	34,685,444	1,274
9	GCA_039737535.1	Lrho_1.0	*Lagodon rhomboides*	785,333,011	3,525,568	31,415,645	2,704
10	GCA_003309015.1	ASM330901v1	*Sparus aurata*	830,381,975	32,411	20,611,436	34,642
11	GCA_020080195.1	ASM2008019v1	*Acanthopagrus latus*	733,109,041	13,476,833	29,922,296	288
12	GCA_013436515.1	ASM1343651v1	*Acanthopagrus latus*	806,072,371	1,174,323	30,688,393	4,265

Table 1. Chromosome level genomes and their metrics for the fishes belonging to the family Sparidae.

Methods

Specimen collection and identification. *Rhabdosargus sarba* was collected from the ICAR – Central Institute of Brackishwater Aquaculture, Muthukadu Experimental Station (Chennai, India) for sequence data generation and genome assembly. The species identity of *R. sarba* was confirmed by amplifying a partial sequence of the Cytochrome C Oxidase I (COI) gene using primers F2 (5′TCGACTAATCACAAAGACATCGGCAC3′) and R1 (5′TAGACTTCTGGGTGGCCAAAGAATCA3′)[9]. The COI gene sequence was submitted to GenBank with the accession number PP779780.

Genome size estimation by flowcytometry. Blood samples for genome size estimation were drawn from the caudal vein of the fish and collected in an anticoagulant solution (38 mM Citric acid, 76 mM Sodium citrate, and 122 mM D-glucose). The blood was centrifuged at 300xg for 5 minutes at 4 °C, and the supernatant was discarded. The pellet was resuspended in 200 µl of 70% ethanol, added drop by drop, and incubated on ice for 30 minutes to fix the cells. After another centrifugation at 300xg for 5 minutes at 4 °C, the supernatant was discarded again. The pellet was then resuspended in 400 µl of phosphate-buffered saline. The resuspended cells (150 µl) were mixed with 600 µl of LB01 buffer (15 mM Tris, 2 mM disodium-EDTA, 0.5 mM spermine tetrahydrochloride, 80 mM KCl, 20 mM NaCl, and 0.1% v/v TritonX-100). The homogenized samples were filtered through a 40 µm sterile cell strainer (Corning Inc., USA) into a 50 ml Falcon tube (Tarson, India). Approximately 500 µl of the filtrate was transferred to a fresh microcentrifuge tube and treated with 1 µl of RNase A (Qiagen, Germany). Propidium iodide (ThermoFisher Scientific, USA) (12 µl) was added to the filtrate for staining the nucleic acid content of the blood sample. For control, 20 µl of chicken erythrocytes from the BD™ DNA QC Particles kit (BD Biosciences, USA) was mixed with LB01 buffer (480 µl), RNase A (1 µl), and propidium iodide (12 µl). The tubes were incubated in the dark at 4 °C for 30 minutes before flow cytometry analysis. Flow cytometry readings were acquired on a BD Accuri™ C6 flow cytometer (BD Biosciences, USA) for a minimum of 10,000 events using a slow flow rate (14 µl/min) and a core size setting of 10 µm. The readings were recorded separately for the control chicken erythrocytes and the sample. The sample was then mixed with control chicken erythrocytes, and the readings for the mixture were taken to assess the genome size. The histogram data was acquired on BD Accuri™ C6 Plus software v1.0.23.1 (BD Biosciences, USA) for genome size estimation. The genome size of *R. sarba* was estimated to be 792.11 Mb based on the flow cytometry data[10,11] (Fig. 2a).

Short read DNA sequence data and genome size prediction. Genomic DNA was extracted from the muscle tissue of *R. sarba* using the DNeasy Blood and Tissue Kit (Qiagen, CA). The quantity of DNA was measured using both the Nanodrop 2000 and Qubit 3 Fluorometer (Thermofisher Scientific, Massachusetts, USA). The integrity

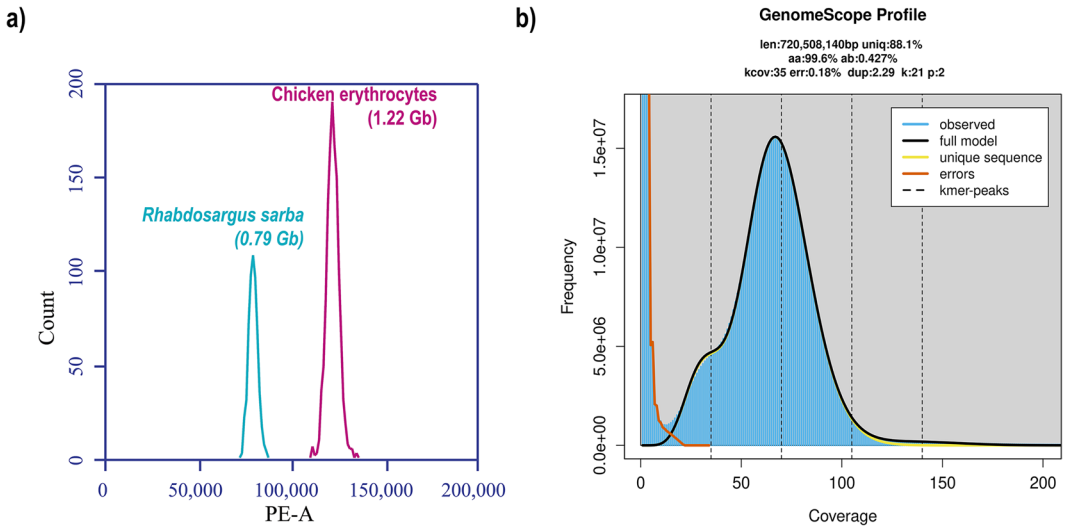

Fig. 2 The genome size estimation profiles of *Rhabdosargus sarba* (**a**) Flowcytometry graph depicting the fluorescent peaks of *R. sarba* and Chicken erythrocytes, (**b**) K-mer based analysis profile for genome size estimation of *R. sarba* using Jellyfish and Genomescope.

	Sequence data (library 1)	Sequence data (library 2)	Sequence data (library 3)
Polymerase Read Bases	404,151,347,955	506,128,555,402	487,071,138,168
Polymerase Reads	5,377,351	5,700,086	5,859,334
Polymerase Read Length (Mean)	75,158	88,793	83,127
Polymerase Read N50	202,518	200,059	196,060
Subread Length (Mean)	8,566	9,877	10,556
Subread N50	11,402	12,737	13,398
Longest Subread Length (Mean)	15,777	18,935	19,073
Longest Subread N50	27,074	30,011	30,692
Unique Molecular Yield	74,421,354,496	95,779,397,632	99,296,886,784
HiFi Reads	1,698,914	2,053,054	1,938,037
Total HiFi bases	18,308,258,214	24,883,335,130	23,527,390,176

Table 2. Summary of the Pacbio sequencing data generated for *R. sarba* from 3 SMRT cells.

of the DNA was assessed on a 1% agarose gel. A library with average size of 328 bp was prepared using Kapa Hyper plus kit (Roche, Basel, Switzerland) and sequenced on Novaseq 6000 platform (Illumina USA) using 2 × 150 bp paired-end chemistry generating approximately 104.3 Gb of data with 91.69% of bases achieving Q30 quality.

Genome size estimation of *R. sarba* was also performed based on short Illumina sequencing reads. It involved counting canonical 21-mers from the paired-end reads with Jellyfish 2.2.3[12] and visualizing the count histograms with Genomescope[13]. This analysis revealed the genome size of *R. sarba* to be 720.50 Mb (Fig. 2b).

Pacbio Long read HiFi Sequence. High molecular weight genomic DNA was extracted from three separate samples of muscle tissues from a single *R. sarba* specimen using Blood and Cell Culture Midi Kit (Qiagen, CA). The quantity was measured with a Qubit 4.0 fluorometer using the DNA HS assay kit (Thermofisher Scientific, USA). DNA purity was assessed using a NanoDrop 2000 spectrophotometer (Thermofisher Scientific, USA) and integrity was confirmed via 1% agarose gel electrophoresis and size estimation using the Agilent FEMTO pulse analyser (Agilent Technologies, USA). Approximately 11.6 µg, 10.2 µg, and 25.6 µg of total DNA were obtained from the three muscle tissue samples. The isolated DNA was sheared into 20 kb fragments using the Megaruptor 3 system (Diagenode, Belgium) to prepare three SMRTbell libraries. Three separate libraries were prepared using the SMRTbell Express Template Preparation Kit 3.0 (Pacific Biosciences, California, USA). These libraries with 17837 bp, 17595 bp, 23744 bp average size underwent purification with AMPure PB beads (Pacific Biosciences, USA) to remove fragments smaller than 5 kb. The purified size-selected libraries were subjected to primer annealing, polymerase binding using the Sequel II Binding Kit 3.2 (Pacific Biosciences, California, USA), and loaded onto separate SMRT cells containing 8 M zero-mode waveguides (ZMWs) for sequencing on the PacBio Sequel II system in CCS/HiFi mode (Nucleome Informatics Pvt Ltd, India). The three libraries yielded 404 Gb, 487 Gb, 506 Gb of polymerase read bases, with a distinct molecular yield accounting for 74 Gb, 99 Gb, 95 Gb (Table 2). Raw polymerase reads were processed using the CCS algorithm v6.4.0 (–min-passes = 3;–min-snr = 2.5;–min-rq = 0.99) to generate HiFi reads.

Attributes	Library statistics
Total Read	356,610,336
Total Bases	53,491,550,400
Raw data in Gb	53.4915504
high-quality (HQ) read pairs (RPs)	49.31%
Clustering usable HQ reads per contig (CTGs > 5KB)	356,590.71
RPs > 10KB apart	33.12%
RPs > 10KB apart (CTGs > 10KB)	65.76%
Intercontig RPs on CTGs > 10KB	48.27%
Intercontig HQ RPs	45.42%
Same strand RPs	43.61%
Split reads	41.91%

Table 3. Summary of the metrics of HiC sequence data.

Sample	Reads (number)	Bases	Q20 bases (%)	Q30 bases (%)	GC (%)
Muscle	55,866,930	8,446,879,776	97.98	94.55	48.20
Spleen	53,934,112	7,881,787,302	95.49	90.63	41.70
Stomach	67,489,682	10,493,325,068	98.24	94.88	46.82
Heart	56,913,844	8,465,588,748	95.94	91.26	44.32
Gills	64,160,926	10,005,128,396	97.64	93.78	44.53
Eyes	65,200,944	9,925,325,492	97.80	94.15	45.33
Kidney	55,435,982	8,412,468,648	96.26	91.77	43.33
Liver	65,601,346	9,686,940,512	98.18	94.92	48.12
Brain	63,196,746	9,762,482,252	96.90	92.62	41.57
Skin	41,214,366	6,393,246,218	97.98	94.50	48.12
Gonad	75,254,790	11,357,362,198	98.28	95.01	47.49

Table 4. Basic stats of RNASeq data generated from different tissues of *R. sarba*.

Hi-C reads. Tissue cross linking, HiC library preparation and amplification was carried out using the Proximo HiC Kit, animal (Phase Genomics, USA) as per manufacturer's recommendations. The restrictions enzymes DpnII, DdeI, HinfI, and MseI were used to prepare HiC library. Quality control of the library was conducted using a Bioanalyzer 2100 (Agilent Technologies, California, USA). A library passing QC was then loaded onto an S4 flow cell at approximately 10 nM concentration and sequenced by the Novaseq 6000 system (Illumina, USA) using 2 × 150 bp paired-end chemistry. This sequencing effort resulted in generating approximately 53.49 Gb of data from 356.6 million reads (Table 3). Hi-C reads were pre-processed with the fastp v0.12.4[14] tool to remove adaptors and perform quality trimming.

RNASeq data generation. Eleven tissues (muscle, spleen, stomach, heart, gills, eyes, kidney, liver, brain, skin, gonad) collected from adult *R. sarba* fish were flash frozen in liquid nitrogen and stored at −80 °C. The total RNA was extracted using TRIzol (DSS Takara, CA, USA) method and purified using Nucleospin MN RNA cleanup kit (Machery Nagel, Germany). RNA quantification was carried out using Qubit 3.0 fluorometer using RNA HS kit (Thermofisher Scientific, Massachusetts, USA) and Nanodrop 2000 (Thermofisher Scientific, Massachusetts, USA). RNA integrity was checked using Bioanalyzer 2100 (Agilent Technologies, California, USA). The cDNA synthesis and library preparation for the eleven samples were prepared with KAPA HyperPrep kit (Roche, Basel, Switzerland). The libraries were subjected to fragment analysis using Bioanalyzer (Agilent Technologies, California, USA) following manufacturer's protocol to evaluate the insert sizes.

Sequencing data was obtained from the qualified libraries using Illumina NovaSeq 6000, S4 Flow Cell (2x150bp read lengths) (Table 4). The RNAseq data were quality trimmed using trimmomatic v0.3911[15].

Genome Assembly of *Rhabdosargus sarba*. The HiFi reads from the three libraries were screened for adaptor and foreign DNA contamination using NCBIs foreign contamination screen tool[16]. The clean reads from the three libraries were combined to obtain about 87.2x coverage of HiFi reads. Hifiasm v0.19.9-r616 tool[17] was used to assemble the clean reads to obtain a contig level assembly. The assembly was about 768 Mb long comprising of 112 contigs and an N50 Value of 30.5 Mb. The contig level assembly was subjected to removal of duplicates using purge-dups tool[18]. [16]The Hi-C reads were pre-processed using fastp v0.12.4 tool[14] for removal of adaptor and quality trimming before utilising them for scaffolding of the contig level assemblies. The scaffolding was performed using the YaHS v1.1 tool[19] and Arima Genomics' mapping pipeline. The primary assembly was again screened for removal of contaminating scaffold sequences using NCBI's Foreign contamination screen tool[16]. The final primary assembly consisted of 31 scaffolds representing the nuclear genome with total length of 764.59 Mb

	Contigs	Scaffolds
Number	112	31
Total Length, bp	768,860,290	764,592,746
Longest sequence, bp	40,195,359	40,195,359
N50, bp	30,587,856	33,989,030
L50	12	11
GC, %	41.77%	41.73%

Table 5. Genome assembly statistics.

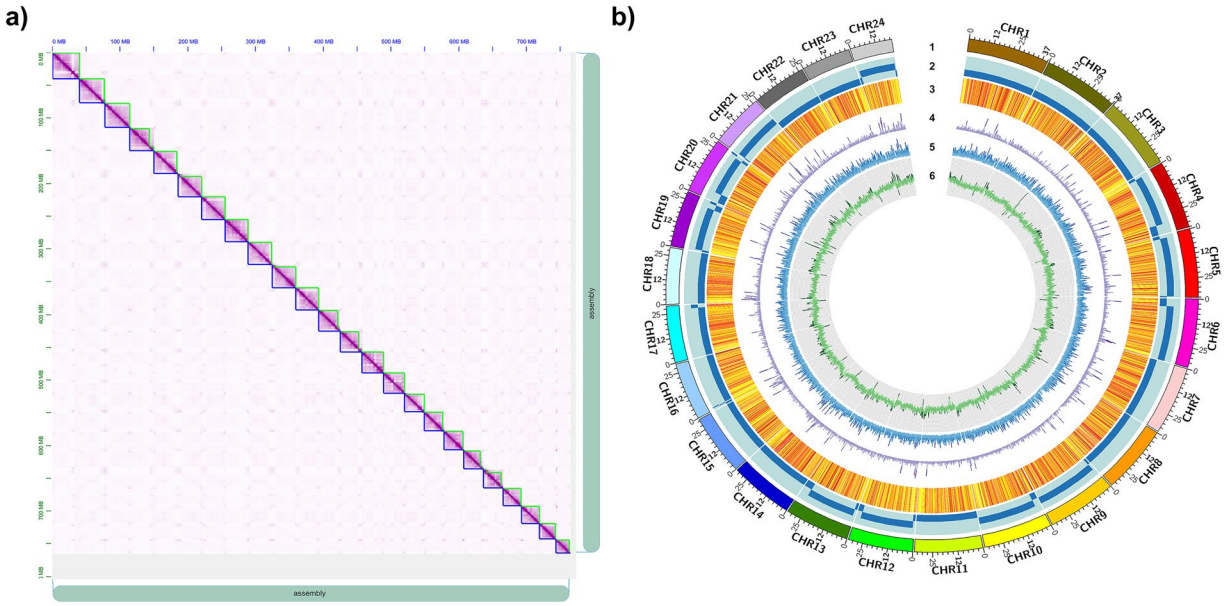

Fig. 3 (**a**) Hi-C contact map representing the assembled genome of *R. sarba* into 24 chromosome level scaffolds. The blue line and the green line indicate scaffold borders and contig borders respectively. (**b**) Circos plot of the Goldlined seabream genome. From the outermost: Track1: The 24 chromosomes of the seabream genome. Track2: Contigs corresponding to the 24 chromosomes represented as tiles. Track3: The protein encoding genes shown as heatmap with progressive colors (yellow–orange–red) indicating length of genes. Track4: Full length isoform sequences supporting the genes of seabream shown as line diagram with incremental isoform lengths shown from purple to dark purple. Track5: Transcripts generated from RNAseq data supporting the genes of seabream shown as line diagram with incremental transcripts lengths shown from blue to dark blue. Track6: GC content of seabream genome shown as line diagram plotted with 50 kb sliding window. The GC values below 35 and above 45 are shown in dark green color, and remaining in green color.

with an N50 value of 33.9 Mb (Table 5), and 1 scaffold representing the mitochondrial genome of length 16,662 bp. The scaffold representing the mitochondrial genome was identified by performing a BLAST[20] analysis using the mitochondrial genome sequence of *R. sarba* (NCBI Accession: KM272585.1) as query against the assembled genome. The top 24 scaffolds covered 99.96% of the assembled genome indicating a high-quality chromosome level assembly as *R. sarba* is reported to contain 24 number of haploid chromosomes[21] (Fig. 3).

Repeat prediction and Masking. *De novo* repeat prediction was performed using Repeat Modeler v2.0.5[22] by employing the rmblast v2.14.1(https://www.repeatmasker.org/rmblast/) search engine, a repeat modeler compatible modified version of NCBI's blastn[20] program, and also enabling the LTR structural analysis. This generated a *de novo* repeat library for the *R. sarba* species, and was used for identifying the repeat profile and masking it with the help of Repeatmasker v4.1.5[23], before proceeding for the gene prediction step. The repeat profile of *R. sarba* is given in Table 6.

Gene prediction and annotation. Genes in *R. sarba* were predicted by combining evidences from ab initio predictions, Illumina RNA-seq datasets and Pacbio Iso-seq datasets. The methodology described earlier in Katneni *et al.*[24] was followed with minor changes. Briefly, *ab initio* predictions were made with AUGUSTUS v3.4.0[25] by giving repeat masked genome as input. Long isoform sequence data for ten tissues of goldlined Seabream (SRA accessions, SRR31204432 – SRR3120441) comprising of full-length non concatemer (FLNC) read data was downloaded from the SRA Database of NCBI. The downloaded bam files were processed by "*cluster*" module of isoseq3 v3.8.2 from pacbio suite using default parameters. The resulting high quality clustered bam

Sequences	31		
Total length	764,592,746 bp		
GC level	41.73%		
Bases masked	219,549,016 bp (28.71%)		
Repeat type	**Number of elements**	**Length occupied, bp**	**Percentage of sequence, %**
Retroelements	117,451	36,487,120	4.77
SINEs:	16,844	2,133,514	0.28
LINEs:	67,123	18,495,992	2.42
L2/CR1/Rex	42,396	9,082,497	1.19
R1/LOA/Jockey	3,702	896,835	0.12
R2/R4/NeSL	1,182	800,204	0.1
RTE/Bov-B	6,914	3,583,896	0.47
L1/CIN4	5,983	2,146,160	0.28
LTR elements:	33,484	15,857,614	2.07
BEL/Pao	2,710	1,631,581	0.21
Ty1/Copia	79	86,283	0.01
Gypsy/DIRS1	18,780	9,285,398	1.21
Retroviral	3,142	2,229,141	0.29
DNA transposons	222,685	41,832,677	5.47
hobo-Activator	96,665	17,618,214	2.3
Tc1-IS630-Pogo	31,976	7,529,899	0.98
MULE-MuDR	228	98,680	0.01
PiggyBac	3,670	818,927	0.11
Tourist/Harbinger	15,825	3,965,976	0.52
Other (Mirage, P-element, Transib)	2,530	433,248	0.06
Rolling-circles	14,634	7116,483	0.93
Unclassified:	747,530	112,776,739	14.75
Total interspersed repeats:		191,096,536	24.99
Small RNA:	7,900	1,028,204	0.13
Satellites:	2,166	254,670	0.03
Simple repeats:	378,454	18,649,435	2.44
Low complexity:	40,597	2,369,568	0.31

Table 6. Repeat profile of *Rhabdosargus sarba* genome assembly.

Tissue	Number of Transcripts	Total size (bp)	longest transcript (bp)	N50 (bp)	SRA Accession number
Brain	48,920	82,274,592	6,932	1,852	SRR31204434
Eyes	30,553	65,050,593	8,174	2,297	SRR31204437
Gills	32,764	60,617,334	5,161	1,982	SRR31204438
Gonad	48,128	84,350,218	4,857	1,912	SRR31204432
Heart	22,652	37,470,971	6,131	1,824	SRR31204439
Kidney	24,387	30,367,465	5,004	1,320	SRR31204436
Liver	14,756	30,523,740	7,922	2,276	SRR31204435
Muscle	12,936	27,258,301	6,123	2,303	SRR31204441
Skin	11,864	23,011,189	7,640	2,143	SRR31204433
Stomach	1,195	1,252,286	2,336	1,142	SRR31204440

Table 7. IsoSeq data statistics used for gene prediction.

file was converted to fasta format using bam2fasta v3.1.1 to obtain the final high-quality full-length transcripts (Table 7). The full-length transcripts were aligned to the genome using GMAP v2021-12-17[26]. The alignments were further used to generate hints and given as input to Augustus to produce Iso-seq based gene predictions. Evidence for gene structures were also generated from RNA-seq data of 11 tissues (SRA accessions, SRR31204118 - 128) by aligning the reads to the genome using Hisat v2.2.1[27], transcript assembly using Stringtie v2.2.1[28] and further processed using TransDecoder v5.7.0[29]. Finally, EVidenceModeler v2.0[29] was used to combine all the evidences in a weighted manner with AUGUSTUS *ab-initio*, Augustus with Iso-seq hints, Iso-seq mapping with GMAP, and RNASeq data having weights 3,9,5 and 4 respectively, to generate a consensus gene models for *R. sarba* genome. A total of 26,913 genes were predicted, out of which 21,274 were found to be complete with start and stop codon and 5,640 were found to be incomplete with either start or stop codon missing (Table 8). The

Property		Value
All	Count	26,913
	Transcripts Per Gene	1
	Average Length	13,240.56
	Median Length	7,033
	Total Length	356,356,492
	Average Coding Length	1,511.84
	Median Coding Length	1,086
	Total Coding Length	40,689,759
	Total Exons	236,873
Complete	Count	21,273
	Average Length	10,737.98
	Median Length	6,240
	Total Length	228,439,756
	Average Coding Length	1,545.72
	Median Coding Length	1,161
	Total Coding Length	32,883,735
	Total Exons	1,88518
Incomplete	Count	5,640
	Average Length	22,680.27
	Median Length	14,736
	Total Length	127,916,736
	Average Coding Length	1,384.05
	Median Coding Length	648
	Total Coding Length	7,806,024
	Total Exons	48,355

Table 8. The protein encoding genes predicted in the Goldlined seabream genome.

Attributes	No. of Transcripts	Percentage
Predicted Genes	26,913	100.00%
Genes with InterProScan Hits	21,611	80.29%
Genes with Blast Hits	21,600	80.25%
Genes with EggNOG annotation	14,904	55.38%
Genes with GO Annotation	19,607	72.85%
Genes with GO Mapping	6,829	25.37%
Genes with Enzyme code	6,541	24.30%
Annotated genes (overall)	22,856	84.92%

Table 9. Annotation statistics of predicted protein encoding genes.

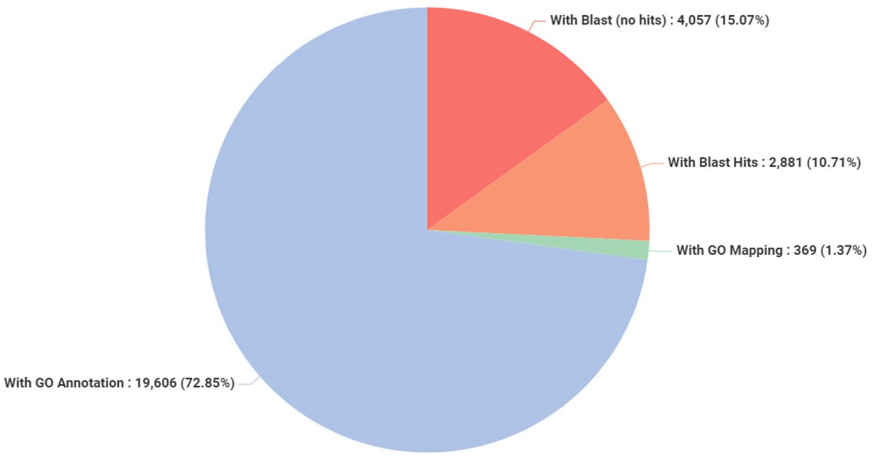

Fig. 4 The functional annotation levels using Omicsbox tool. Pie chart depicting the percentage of transcripts with GO annotation, with only GO mapping, with only blast hits and No hits.

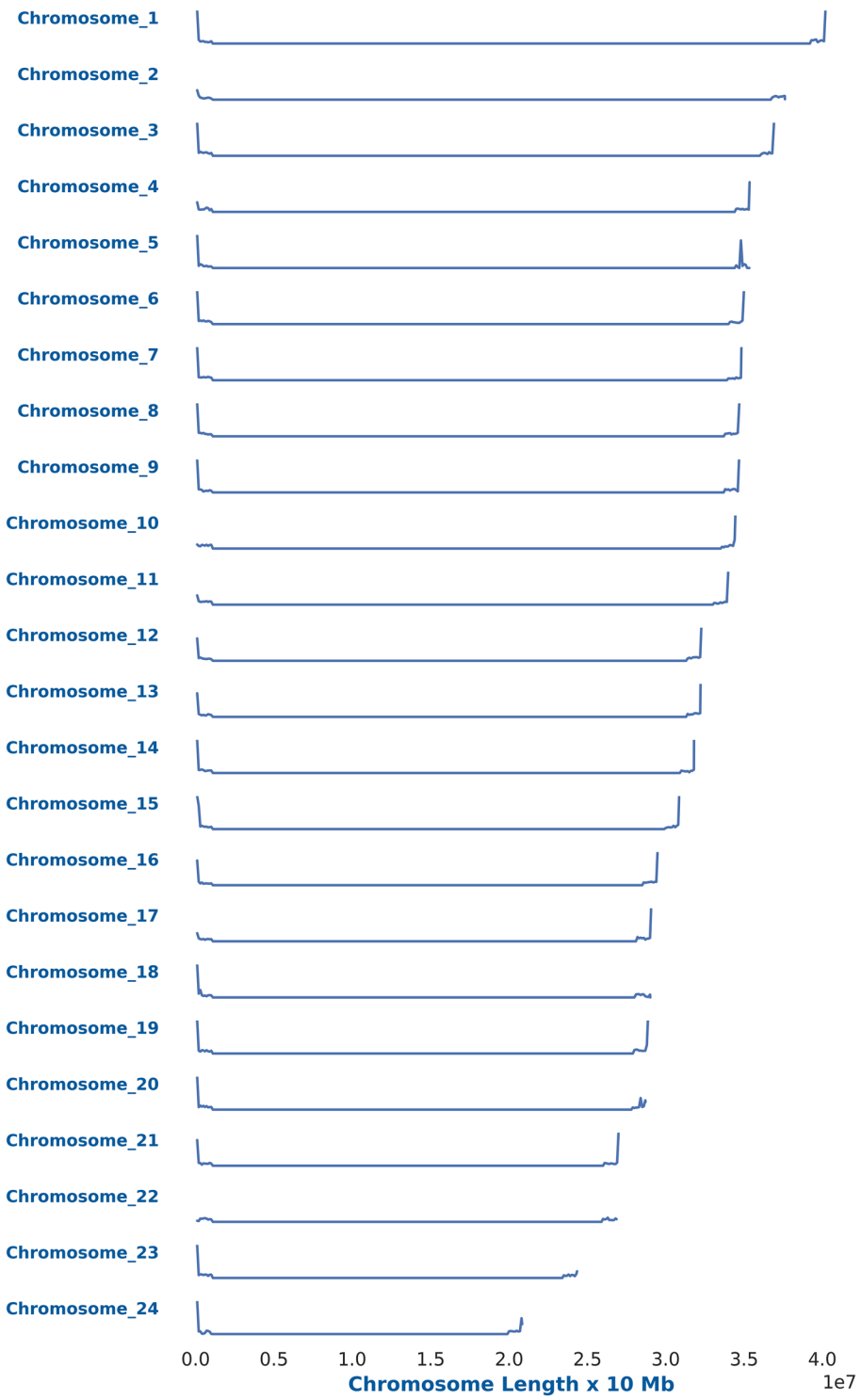

Fig. 5 The graph depicting the telomeric repeats present at the ends of the chromosomes assembled in the *R. sarba* genome.

annotation of Protein encoding genes (PEGs) was conducted using a combination of results from multiple tools. The blastx[20] tool was used against the Actinopterygii (txid7898) dataset from NCBI's non-redundant database. Additionally, InterProScan[30] was employed for annotation based on identifying protein domains, functional motifs, and assigning Gene Ontology (GO) terms through protein signature recognition. EggNOG[31] mapper provided orthology-based functional annotation and pathway prediction. The OmicsBox Tool v3.0.25[32] was used to combine the results from all the three tools to obtain functional annotation (as shown in Table 9 and Fig. 4). The pathway analysis was performed by mapping to KEGG Database[33] revealing that the annotated PEGs were participating in 332 pathways (List of pathways, https://doi.org/10.6084/m9.figshare.28092089.v2[34]).

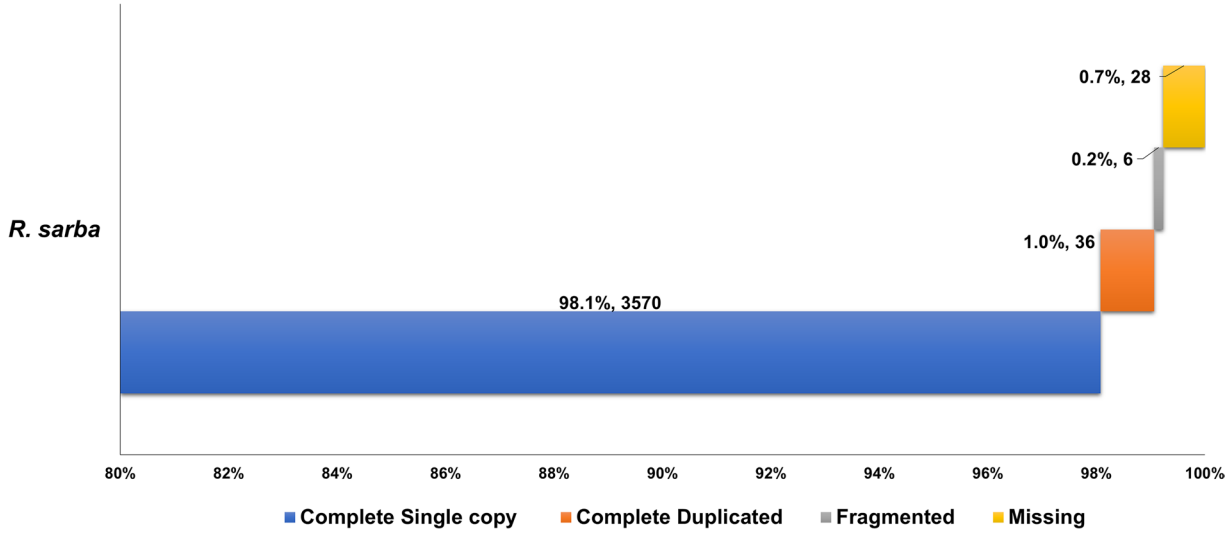

Fig. 6 Graph depicting BUSCO analysis for estimating genome completeness of *R. sarba* against Actinopterygii_odb10 orthologous database.

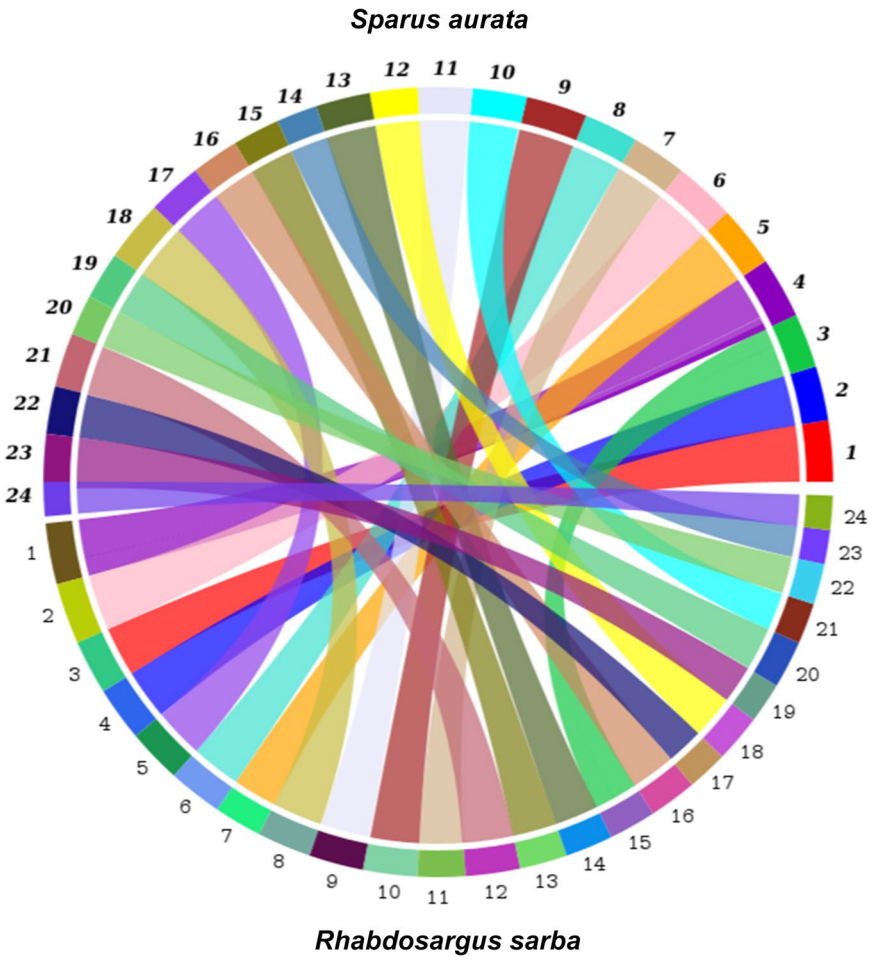

Fig. 7 Synteny map of *R. sarba* (goldlined seabream) with *Sparus aurata* (gilthead seabream).

Data Records

The raw datasets including PacBio-Hifi Reads (SRA Accession: SRR31203107 - SRR31203109), Illumina short DNA reads (SRA Accession: SRR31203208), Illumina Hi-C reads (SRA Accession: SRR3120881), Illumina RNASeq reads (SRA Accession: SRR31204118 - SRR31204128) and PacBio IsoSeq FLNC reads

S. No.	Data	Alignment rate (%)	Biosample Accession number	SRA Accession numbers
1	DNA - Illumina short reads	97.99	SAMN44551992	SRR31203208
2	RNA - Muscle	95.28	SAMN44551992	SRR31204128
3	RNA - Spleen	89.05	SAMN44553731	SRR31204127
4	RNA - Stomach	93.69	SAMN44553732	SRR31204125
5	RNA - Heart	77.74	SAMN44553733	SRR31204124
6	RNA - Gills	93.51	SAMN44553734	SRR31204123
7	RNA - Eyes	92.44	SAMN44553735	SRR31204122
8	RNA - Kidney	85.19	SAMN44553736	SRR31204121
9	RNA - Liver	97.19	SAMN44553737	SRR31204120
10	RNA - Brain	91.55	SAMN44553738	SRR31204119
11	RNA - Skin	94.92	SAMN44553739	SRR31204118
12	RNA - Gonad	95.95	SAMN44553740	SRR31204126

Table 10. Read alignment of Illumina short read DNA and RNA datasets generated in this study against the *R. sarba* genome.

(SRA Accession: SRR31204432 - SRR31204441) were deposited under the BioProject accession number PRJNA1180767[35] at NCBI with the SRA study accession number SRP542769[36]. The assembly has been uploaded to NCBI Genbank's Genome assembly database with accession number GCA_047275855.1[37]. The genome assembly and predicted genes are uploaded to figshare repository with https://doi.org/10.6084/m9.figshare.28092089.v4[34].

Technical Validation

The genome assembly contiguity and quality were assessed based on 9 parameters. (1) the assembly stats like N50 length and L50 value, (2) the number of chromosome level contigs (3) percentage of the assembled genome in the top longest scaffolds equivalent to the chromosome number of *R. sarba*, (4) the presence of full-length mitochondrial DNA in a single scaffold, (5) the presence of telomeric repeats in the assembled scaffolds, (6) the genome completeness assessed with BUSCO v5.7.0 tool[38], (7) the quality value (QV) and error rate in the genome calculated by using Merqury v1.3 tool[39] (8) Synteny with other related species with chromosome level assembly, and (9) read representation stats.

The highly contiguous assembly at contig level consisted of 112 contigs with N50 length of 30.58 Mb and L50 value of 12. The scaffold level assembly consisted of 31 scaffolds with an N50 length of 33.98 Mb and L50 value of 11. Also 11 of the contigs were found to be at chromosome level, indicating a reference level assembly considering a chromosome number of 24 for *R. sarba*[21]. The assembled genome indicated a chromosome level assembly with about 99.96% of the assembly size encompassed by the first 24 scaffolds. This indicates that the 7 unplaced scaffolds and the mitochondrial genome sums upto only <0.04% (243.37 Kb) of the assembly size, having an average length of just 30.42 Kb. The full-length mitochondrial genome was observed in a single scaffold of 16,662 bp. This reflects the contiguity of the assembly. The mitochondrial genome obtained, was further annotated by the mitoannotator module of MitoFish database v4.09[40] (Fig. 5). The annotation revealed the presence of 13 protein codings genes, 22 tRNAs, 2 rRNAs and a D-loop, which are consistent with the mitochondrial genome structure. The chromosomes were further scanned for the presence of telomeres using tidk v0.2.0 tool[41] with the search module of the tool for the telomeric repeat sequence (AACCCT) and default parameters. All the chromosomes were observed to contain telomeric ends (Fig. 5). The completeness assessment using BUSCO v5.7.0 tool[38] against actinopterygii_odb10 (2021-02-19) ortholog database, revealed a highly complete genome of 99.3%, consisting of 99.1% (3570) of complete orthologs and 0.2% (6) of fragmented orthologs and only 0.7% (28) of missing orthologs (Fig. 6). The quality value and error rate evaluated by Merqury v1.3 tool[39] was 62.86 and 5.16607e-07 respectively. This implies the possibility of only 1 error base in 1.9×10^6 bases indicating high base accuracy of genome assembly. Synteny analysis was conducted on the *R. sarba* in comparison with the gilthead seabream (*Sparus aurata*) [NCBI RefSeq assembly accession: GCF_900880675.1] using SyMAP v5.3.5[42]. The 24 chromosomes of Gilthead seabream comprise of length 818,505,427 bp were aligned to the 24 chromosomes of *R. sarba* comprising of length 764,366,029 bp to assess the similarity between the two. It was observed there were 26 syntenic blocks with length greater than 10 Mb and no blocks were found with lengths less than 100 Kb between the two genomes indicating a very high co-relation between them. The coverage statistics indicate that both the genomes share about 98% similarity (Fig. 7). The assembly was further validated by the alignment percentage obtained from aligning RNA-seq reads and DNA short reads onto the genome (Table 10).

Code availability

All data processing commands and pipelines were carried out in accordance with the instructions and guidelines provided by the relevant bioinformatics software. There were no custom scripts or code utilized in this study.

Received: 30 December 2024; Accepted: 4 June 2025;
Published online: 11 July 2025

References

1. Yang, J. *et al*. Whole-Genome Resequencing Reveals Signatures of Adaptive Evolution in *Acanthopagrus latus* and *Rhabdosargus sarba*. *Animals* **14**, 2339 (2024).
2. Jiang ShiGui, J. S., Liu HongYan, L. H., Su TianFeng, S. T. & Gong ShiYuan, G. S. Molecular phylogeny of mitochondrial cytochrome b gene sequences from four Sparidae fishes. *Journal of Fishery Sciences of China* **10**, 184–188 (2003).
3. Li, J. *et al*. The complete mitochondrial genome of the *Rhabdosargus sarba* (Perciformes: Sparidae). *Mitochondrial DNA Part A* **27**, 1606–1607 (2016).
4. Katneni, V. K. *et al*. Genome assembly at chromosome scale with telomere ends for Pearlspot, *Etroplus suratensis*. *Sci Data* **11**, 1226 (2024).
5. Vij, S. *et al*. Chromosomal-level assembly of the Asian seabass genome using long sequence reads and multi-layered scaffolding. *PLoS Genet* **12**, e1005954 (2016).
6. Shekhar, M. S. *et al*. First Report of Chromosome-Level Genome Assembly for Flathead Grey Mullet, *Mugil cephalus* (Linnaeus, 1758). *Front Genet* **13**, 911446 (2022).
7. Shekhar, M. S. *et al*. Genome assembly, Full-length transcriptome, and isoform diversity of Red Snapper, *Lutjanus argentimaculatus*. *Sci Data* **11**, 796 (2024).
8. Zhao, N., Guo, H.-B., Jia, L., Dong, Z. & Zhang, B. The genome assembly of flathead grey mullet *Mugil cephalus*. *Front Mar Sci* **9**, 988397 (2022).
9. Ward, R. D., Zemlak, T. S., Innes, B. H., Last, P. R. & Hebert, P. D. N. DNA barcoding Australia's fish species. *Philosophical Transactions of the Royal Society B: Biological Sciences* **360**, 1847–1857 (2005).
10. Swathi, A., Shekhar, M. S., Katneni, V. K. & Vijayan, K. K. Genome size estimation of brackishwater fishes and penaeid shrimps by flow cytometry. *Mol Biol Rep* **45**, 951–960 (2018).
11. Raymond, J. A. J. *et al*. Comparative genome size estimation of different life stages of grey mullet, *Mugil cephalus* Linnaeus, 1758 by flow cytometry. *Aquac Res* **53**, 1151–1158 (2022).
12. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
13. Vurture, G. W. *et al*. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204 (2017).
14. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
15. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
16. Astashyn, A. *et al*. Rapid and sensitive detection of genome contamination at scale with FCS-GX. *Genome Biol* **25** (2024).
17. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**, 170–175 (2021).
18. Guan, D. *et al*. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).
19. Zhou, C., McCarthy, S. A. & Durbin, R. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics* **39** (2023).
20. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403–410 (1990).
21. Arkhipchuk, V. V. Chromosome database. Database of Dr. Victor Arkhipchuk. (1999).
22. 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).
23. Chen, N. Using Repeat Masker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* **5**, 4–10 (2004).
24. Katneni, V. K. *et al*. A Superior Contiguous Whole Genome Assembly for Shrimp (*Penaeus indicus*). *Front Mar Sci* **8** (2022).
25. Hoff, K. J. & Stanke, M. Predicting Genes in Single Genomes with AUGUSTUS. *Curr Protoc Bioinformatics* **65**, 1–54 (2019).
26. Wu, T. D. & Watanabe, C. K. GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
27. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907–915 (2019).
28. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290–295 (2015).
29. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9**, 1–22 (2008).
30. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
31. Huerta-Cepas, J. *et al*. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol Biol Evol* **34**, 2115–2122 (2017).
32. Omicsbox. OmicsBox-Bioinformatics made easy (Version 3.0.25). (2019).
33. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* **45**, D353–D361 (2017).
34. Shekhar, M. S. *et al*. Chromosome-level genome assembly with telomeric repeats at scaffold ends for *Rhabdosargus sarba*. https://doi.org/10.6084/m9.figshare.28092089.v4 (2024).
35. *NCBI Bioproject* http://identifiers.org/bioproject:PRJNA1180767 (2024).
36. *NCBI Sequence Read Archive* http://identifiers.org/insdc.sra:SRP542769 (2024).
37. *NCBI Genbank* http://identifiers.org/insdc.gca:GCA_047275855.1 (2025).
38. 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. *Mol Biol Evol* **38**, 4647–4654 (2021).
39. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: Reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol* **21** (2020).
40. Zhu, T., Sato, Y., Sado, T., Miya, M. & Iwasaki, W. MitoFish, MitoAnnotator, and MiFish Pipeline: Updates in 10 Years. *Mol Biol Evol* **40**, msad035 (2023).
41. Brown, M., la Rosa, P. M. & Mark, B. A Telomere Identification Toolkit. *Zenodo* https://doi.org/10.5281/zenodo.10091385 (2023).
42. Soderlund, C., Bomhoff, M. & Nelson, W. M. SyMAP v3. 4: a turnkey synteny system with application to plant genomes. *Nucleic Acids Res* **39**, e68–e68 (2011).

Acknowledgements

This work was carried out in the project entitled 'Genome sequencing and its application for Brackishwater Aquaculture' funded by Department of Biotechnology, Government of India (BT/PR47080/AAQ/3/1050/2022).

Author contributions

M.S.S., V.K.K., A.K.J. and K.K.L. designed the study. M.S.S., V.K.K. and A.K.J. led the research, and wrote the draft manuscript. V.K.K., A.K.J., K.K. and S.K.P. standardized the pipelines for genome assembly and annotation. S.K.P., K.K. and J.R.J.A acquired the samples and contributed to preliminary analysis, and visualization. All authors contributed substantially to the revisions and 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 V.K.K.

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.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025