scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome level genome and full length transcriptome of long whiskers catfish, *Mystus gulio* (Hamilton, 1822)

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

Mystus gulio **(Hamilton, 1822), the long whiskers catfish, belongs to the family Bagridae and inhabits brackishwater. The current study employed a multi-platform sequencing strategy including, PacBio HiFi, Illumina short read and high-throughput chromosome conformation capturing (Hi-C) sequencing technologies to generate raw data and assemble a chromosome level genome. The final assembly is 706.32 Mb with an N50 length of 22.79 Mb and 96.56% of the assembly anchored to 29 chromosomes. Repeat sequences accounted for 32.66% of the genome, and BUSCO analysis revealed a genome completeness of 98.1% against actinopterygii_odb12. The full-length transcriptomes were also generated with PacBio Iso-Sequencing strategy for ten tissues (gills, kidney, liver, muscle, ovary, skin, stomach, arborescent organ, dorsal barbel and gallbladder). The transcriptome data generated was integrally used to predict 23,339 protein encoding genes, comprehensive classification of isoforms and analysis of alternate splicing events across the tissues. The resources generated in the study will give meaningful insights into its nutritional potential, ecological functions for the genetic improvement, molecular breeding and evolutionary studies of** ***M. gulio*****.**

Background & Summary

The *Mystus gulio* (Hamilton, 1822), commonly known as long whiskers catfish, primarily lives in brackishwater habitats (Fig. 1). This fish is mainly distributed in Asian countries bordering the Eastern Indian Ocean ranging from India to Indonesia and Vietnam[1]. This bagrid catfish falls under small indigenous fish species category with potential role to meet food and human nutrition security especially of vitamins and micronutrients in developing countries[2–4].

In the North-East region of the Sunderbans, the world's largest mangrove ecosystem, the population of this species was reported as declining between 1950 to 2000 with a mean decadal decline of 33.56%[5]. Induced breeding of this catfish followed by larval development has been successfully demonstrated in India and Bangladesh[6–8]. However, major constraints limiting the commercial aquaculture of this catfish are the paucity of wild seeds and unavailability of hatchery produced seeds[9]. So far, the complete genome sequence is not available for the Genus, *Mystus*. The availability of high-quality genome and transcriptome of *M. gulio* will provide essential resources to identify genes and molecular pathways associated with aquaculture related traits such as growth, disease resistance, salinity tolerance and other stress responses. This will facilitate the development of genetic markers and quantitative trait loci (QTLs) enabling marker assisted breeding and genetic improvement programs. In addition, these resources will also assist in the characterization of genes and metabolic pathways leading to development of genome scale metabolic models, improving our understanding of feed conversion efficiency and nutritional potential of *M. gulio*. Furthermore, these genomic resources will also enable population genetics studies assessing genetic diversity and adaptive variation across populations inhabiting different natural environments, thereby supporting conservation and stock improvement programs. Overall, these resources lay the groundwork for evidence-based genetic improvement and sustainable management of *M.gulio* in both aquaculture and natural habitats.

ICAR-Central Institute of Brackishwater Aquaculture, 75, Santhome High Road, MRC Nagar, Chennai, 600028, Tamil Nadu, India. ✉e-mail: msshekhar@hotmail.com

Fig. 1 A specimen of long whiskers catfish, *M. gulio*.

In this study, we have used about 57.18 Gb (81X coverage) of PacBio HiFi DNA sequence reads to generate highly contiguous genome assembly with a total length of 706.31 Mb in 283 contigs and N50 length of 14.99 Mb. Thereafter, 91.27 Gb (129X coverage) of Arima-HiC linked reads were used to order and orient the contigs to scaffolds. The final assembly generated with HiFi and Hi-C reads is of 706.32 Mb length in 108 scaffolds with N50 length of 22.79 Mb. The actual genome assembly length is higher than the estimates obtained with flow cytometry method and k-mer histogram profiles but similar to the genome length reported for other bagrid fishes. The finished assembly consisted of 29 chromosome-scale scaffolds similar to the chromosome numbers reported for *M. gulio* and also the genus, *Hemibagrus*[10,11] while other bagrid fishes of the genera, *Tachysurus*[12–15] and *Leiocassis*[16] were reported to contain 26 chromosomes. Overall, 96.56% of the assembly length is represented in 29 chromosome-scale scaffolds. The genome assembly is estimated to have a quality value of 59.49. Telomeric ends were identified at the ends of the chromosome-level scaffolds.

The *de novo* repeat library generated from the current assembly consisted of 2,174 repeat families, which were used to classify the repeat elements in the assembly. Totally 32.66% of the genome was found to be repeat elements. A majority of the repeat elements remained unclassified (13.11%), whereas the classified repeat elements were dominated by simple repeats (6.53%), followed by DNA transposons (5.51%), LINEs (3.44%) and LTR elements (3.20%). The protein encoding genes (PEGs) predicted based on hints from Illumina RNAseq reads, PacBio Iso-Seq reads, and ab initio methods consisted of 23,339 sequences. The high-quality chromosome level genome and transcriptome resources generated in the study would help us in identifying the genetic basis of many of the characteristics of *M. gulio* and evolutionary variations in bagrid catfishes.

Methods

Specimen collection and identification. The *M. gulio* specimen was collected from the Muttukadu Experimental Station of ICAR-Central Institute of Brackishwater Aquaculture, Chennai, India. The partial sequence of Cytochrome C Oxidase I (COI) gene was amplified and sequenced using primers F2 (5′TCGACTAATCACAAAGACATCGGCAC3′) and R1 (5′TAGACTTCTGGGTGGCCAAAGAATCA3′)[17], to confirm the species identity of *M. gulio*. The sequence was submitted to GenBank under the accession number PP855629. The COI sequence was utilized in phylogenetic analysis to ascertain the lineage of the *M. gulio* sample that was used in this study. The various *Mystus* species COI sequence accessions from Barcode of Life Data Systems (BOLD database) were used along with *Aphyocharax anisitsi* serving as an outgroup for the phylogenetic analysis. The total number of accessions utilized to create the phylogenetic tree is listed in Table 1. The Hasegawa-Kishino-Yano model with Gamma and invariant site distribution rates (HKY + G + I), 1000 bootstrap iterations were used for

Species Name	No. of Accessions	Accession Numbers
Aphyocharax anisitsi (Outgroup)	7	FJ749054, HM562855, JN988678, JN988679, JQ667493, KU288969, KU289034
Mystus albolineatus	5	KF824808, KF824809, KF824810, KF824811, KF824812
Mystus armatus	1	MK883814
Mystus atrifasciatus	5	KF824798, KF824799, KF824800, KF824801, KF824802
Mystus bleekeri	25	JN228943, JN628898, JN628928, JX260916, JX260918, JX983370, KF598806, KJ936689, KJ936764, KP939357, KT364779, KT896741, KX266834, MG675622, MH156939, MH197195, MK440705, MK599492, MK599493, MN083144, MN096212, MN520011, MW485075, OM932505, OP575614
Mystus bocourti	3	EU490863, JQ420129, MK448116
Mystus cavasius	18	JX260919, KJ909456, KJ909457, KJ909458, KJ909459, KX946709, KX946711, KX946712, KX946713, KX946714, KX946715, MF601313, MT654646, MT654657, MT928124, MT928125, MW485072, NC030187
Mystus cineraceus	1	KX886803
Mystus gulio	20	FJ384684, KC595988, KF214302, KF511564, KF574789, KF824813, KJ959643, KX455898, KX455905, MF611595, MK359910, MK681762, MK902728, MK902732, MK995086, MN083111, MN458398, MT928126, MT997926, PP855629*
Mystus malabaricus	22	KX073598, KX946716, KX946717, KX946718, KX946719, KX946720, KX946721, KX946722, KX946723, MN255387, MT928129, MT928130, MT928131, MT928132, MT928134, MT928135, MT928136, MT928137, MT997684, MW074288, OP038554, OP297842
Mystus montanus	6	MF591712, MT928133, MT928138, MT928139, MT928140, MT928142
Mystus mysticetus	9	KF805372, KF805373, KF805374, KF805375, MG981076, MK448092, MK448108, MK448181, MK448190
Mystus ngasep	11	KJ909389, KJ909390, KJ909391, KJ909392, MG736522, MG736523, MG736524, MG736525, MG736526, MG736527, MG736528
Mystus nigriceps	15	KU692645, KU692646, KU692647, KU692648, KU692649, KU692650, KU692651, KU692652, KU692653, KU692654, KU692655, KU692656, KU692657, KU692658, MG981078
Mystus oculatus	8	HQ009492, HQ009493, MT928143, MW686932, MW686933, MW686934, MW686935, MW686936
Mystus pelusius	4	KJ553786, KJ553996, OM366048, OM366049
Mystus singaringan	8	JQ289146, KU692659, KU692660, KU692661, KU692662, MK049456, MK448115, MK448162
Mystus vittatus	24	JQ713855, JX260920, JX983385, KF511563, KJ936694, KJ959638, KR809744, KT364780, KU043353, KY769368, KY867665, MK359861, MK359890, MK359926, MK359946, MK359957, MK599496, MK599497, MK599498, MN082647, MN172312, MW485071, MW485076, NC032082
Mystus wolffii	2	MN243470, MN243483

Table 1. List of accessions used for the construction of phylogenetic tree. *Sequence from the specimen used in the current study.

the construction of a Maximum Likelihood tree using MEGA X software[18]. FigTree v1.4.32[19] was used to visualize the final tree. The phylogenetic tree groups the specimen used in the current study with the *M. gulio* clade (Fig. 2).

Sequencing data. *PacBio Long read HiFi Sequence.* High molecular weight genomic DNA was extracted from three separate samples of muscle tissues from a single *M. gulio* specimen using Blood and Cell Culture Midi Kit (Qiagen, Hilden, Germany). DNA quality was checked using 1% agarose gel electrophoresis. The DNA purity and quantity was measured with NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, Massachusetts, USA) and Qubit 4.0 fluorometer using the DNA HS assay kit (Thermo Fisher Scientific, Massachusetts, USA). DNA size estimation was performed using the Agilent FEMTO pulse analyser (Agilent Technologies, California, USA). The isolated DNA was sheared into 20 kb fragments using the Megaruptor 3 system (Diagenode, Belgium) to prepare three SMRTbell libraries. Three separate libraries of average insert sizes 19,225 bp, 17,380 bp, and 19,762 bp were prepared using the SMRTbell Express Template Preparation Kit 3.0 (Pacific Biosciences, California, USA). These libraries underwent purification with AMPure PB beads (Pacific Biosciences, California, 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 372 Gb,403 Gb, 446 Gb of polymerase read bases, with a distinct molecular yield accounting for 83 Gb, 86 Gb, 100 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. Table 2.

Genomic short-read sequencing. The genomic DNA for short-read Illumina paired-end sequencing was extracted using the DNeasy Blood and Tissue kit (Qiagen, Hilden, Germany). The DNA was quantified using Nanodrop 2000(Thermo Fisher Scientific, Massachusetts, USA) and Qubit 3 Fluorometer (Thermo Fisher Scientific, Massachusetts, USA) and library preparation was performed using the Kapa Hyper plus kit (Roche, Basel, Switzerland) with an insert size of 350 bp. Sequencing of the library was done using 2 × 150 bp paired-end

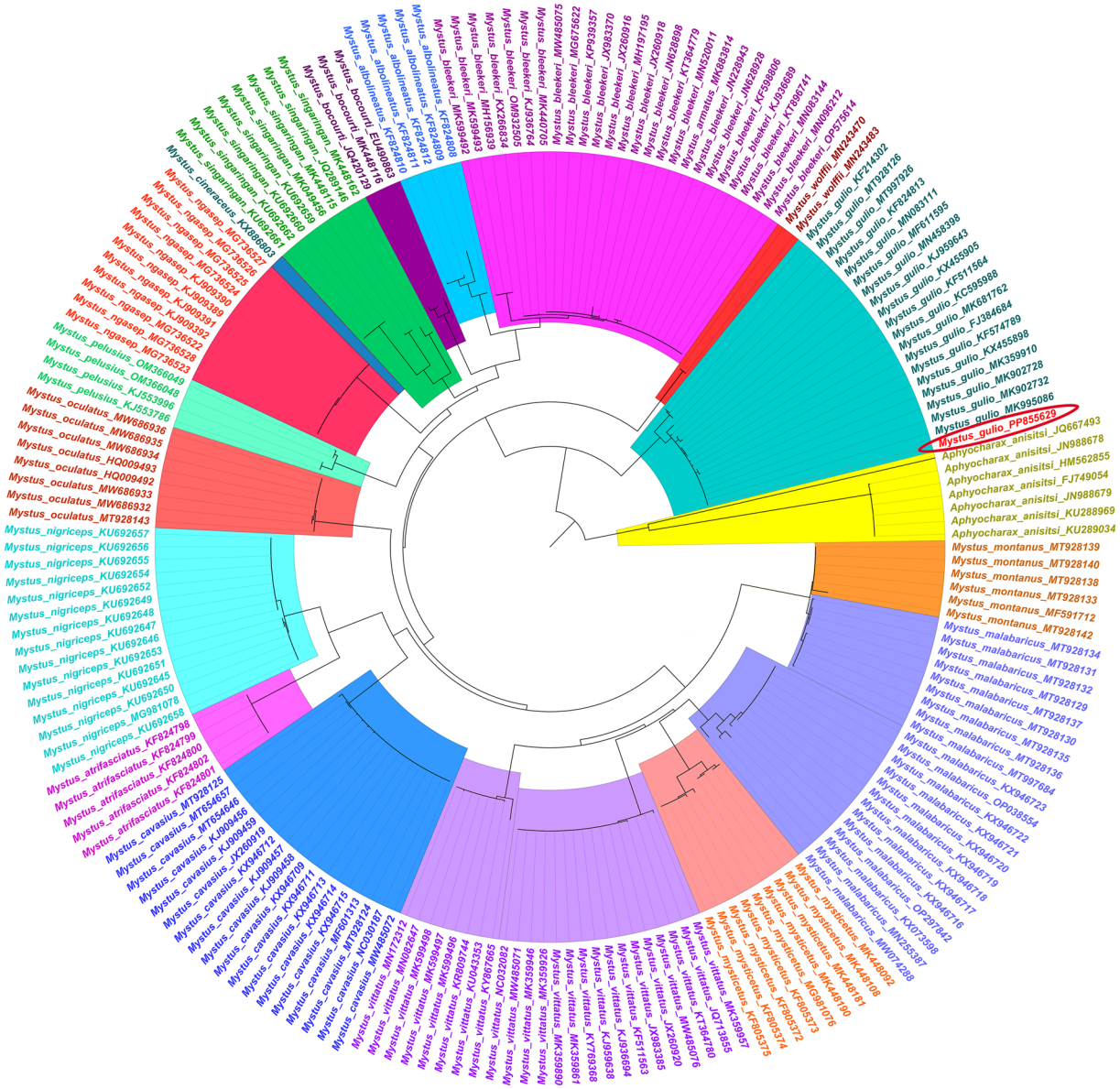

Fig. 2 Maximum likelihood phylogenetic tree based on COI gene sequence of *M. gulio* used in the current study along with sequence of other *Mystus* species and *Aphyocharax anisitsi* as outgroup showing the selected specimen grouping with *M. gulio* clade.

chemistry on the Novaseq 6000 platform (Illumina Inc., California, USA). The data generated consisted of 240 million paired reads summing upto 70.9 Gb (100× coverage).

Hi-C Sequencing. Two Hi-C cross-linking libraries were prepared using the sample from a single adult specimen of *M. gulio*. The first Hi-C library preparation was carried out using the Proximo Hi-C Kit, animal (Phase Genomics, Washington, USA). 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 resulted in generating approximately 45.91 Gb of data from 153 million paired-end reads.

The second Hi-C library was prepared using the EpiTect Hi-C Kit (Qiagen, Hilden, Germany) and quantified using the Qubit 4.0 fluorometer (Thermo Fisher Scientific, Massachusetts, USA) using the DNA HS assay kit (Thermo Fisher Scientific, Massachusetts, USA). The library size was estimated on Agilent Tapestation 4150 (Agilent Technologies, California, USA) utilizing high sensitive D1000 screentapes (Agilent Technologies, California, USA). The libraries were sequenced on the Novaseq 6000 platform (Illumina Inc., California, USA) using 2 × 150 bp paired-end chemistry. This generated 45.36 Gb of data from 142 million paired-end reads.

	Sequence data (library 1)	Sequence data (library 2)	Sequence data (library 3)
Polymerase Read Bases	372,534,886,387	403,841,180,383	446,591,665,701
Polymerase Reads	5,017,547	5,148,377	5,931,532
Polymerase Read Length (Mean)	74,246	78,440	75,291
Polymerase Read N50	188,781	181,135	167,955
Subread Length (Mean)	9,652	10,153	10,601
Subread N50	15,547	15,712	13,697
Longest Subread Length (Mean)	18,687	18,641	18,902
Longest Subread N50	30,414	27,430	28,064
Unique Molecular Yield	83,800,408,064	86,024,454,144	100,047,536,128
HiFi Reads	1,372,338	1,580,752	1,718,927
Total HiFi bases	16,822,116,279	19,957,124,024	20,396,113,579
HiFi Reads N50	19,150	18,264	14,690

Table 2. Summary of the genomic reads from PacBio SMRT cells.

Tissue	Number of FLNC reads	FLNC total length (bp)	Transcripts	Total bases (bp)	Mean transcript length (bp)	Max transcript length (bp)	N50 (bp)	GC %
Arborescent organ	573,321	1,116,731,381	36,150	74,059,758	2,049	7,559	2,197	45.54
Dorsal barbel	342,758	670,994,375	25,667	51,015,464	1,988	6,130	2,162	45.25
gallbladder	545,580	653,322,805	10,805	12,650,388	1,171	3,046	1,276	49.12
Gills	334,077	619,333,916	23,931	46,866,486	1,958	5,206	2,104	47.31
Kidney	207,301	258,158,977	8,264	10,038,447	1,215	3,710	1,416	46.28
Liver	416,519	689,321,852	35,987	62,424,936	1,735	5,553	2,058	46.00
Muscle	498,516	672,942,963	46,983	65,585,846	1,396	4,658	1,518	47.27
Ovary	685,376	1,401,730,275	38,758	80,292,370	2,072	6,031	2,253	47.48
Skin	464,215	875,697,827	31,604	59,602,152	1,886	5,816	2,010	46.35
Stomach	391,104	606,943,521	6,985	10,939,014	1,566	4,110	1,607	48.90

Table 3. Statistics for Iso-Seq raw data.

Iso-Seq data generation. Ten tissues (gills, kidney, liver, muscle, ovary, skin, stomach, arborescent organ, dorsal barbel and gallbladder) were collected from the same adult specimen and flash frozen in liquid nitrogen and stored at −80 °C. The total RNA was isolated using the TRIzol (DSSTakara, CA, USA) method and purified using the Nucleospin MN RNA Cleanup kit (Thermo Fisher Scientific, Massachusetts, USA). Isolated RNA was quantified using Nanodrop 2000 (Thermo Fisher Scientific, Massachusetts, USA) and checked on 1% agarose gel. The integrity of the RNA was assessed using Agilent Bioanalyzer 2100 (Agilent Technologies, California, USA). Full-length cDNA synthesis was done using NEBNext Single cell cDNA synthesis kit (New England Biolabs, Hitchin, UK). Library preparation of the cDNA was done using SMRTbell Express template Preparation Kit 3.0 (Pacific Biosciences, California, USA) as per the manufacturer's protocol and sequenced on the PacBio Sequel II platform. The raw subreads was used to extract the full length transcripts as described earlier[20,21]. Briefly, the *CCS* tool of the PacBio suite (https://github.com/PacificBiosciences/pbbioconda) was used for extracting the HiFi reads from the subreads data and further subjected to demultiplexing and primer removal using the *lima* tool of the PacBio suite (https://github.com/PacificBiosciences/pbbioconda). The final full-length non-chimeric (flnc) reads were generated using the *isoseq refine* tool and clustered using the *isoseq cluster* tool to get full-length transcripts. Table 3.

RNASeq data generation. Tissues, including muscle, spleen, stomach, heart, gills, ovary, eyes, kidney, liver, brain, skin, gall bladder, arborescent organ, pectoral fin, dorsal fin, anal fin, and dorsal barbel, were collected from an adult female specimen of *M. gulio*. In total, seventeen distinct tissues were sampled from the specimen, immediately flash-frozen in liquid nitrogen and stored at −80 °C until further processing. Total RNA from the tissues were extracted and purified using TRIzol (DSS Takara, California, USA) method and Nucleospin MN RNA cleanup kit (Machery Nagel, Düren, Germany), respectively. The quantification of the purified RNA was carried out on Qubit 3.0 fluorometer using RNA HS kit (Thermo Fisher Scientific, Massachusetts, USA) and Nanodrop 2000 (Thermo Fisher Scientific, Massachusetts, USA). The integrity of the purified RNA was assessed on Bioanalyzer 2100 (Agilent Technologies, California, USA). The cDNA synthesis and sequencing libraries were prepared using the KAPA HyperPrep kit (Roche, Basel, Switzerland). The library insert size was estimated using Bioanalyzer (Agilent Technologies, California, USA). The libraries were sequenced on the Illumina Novaseq 6000 system (Illumina Inc., California, USA) using 2 × 150bp paired-end sequencing chemistry. The raw read statistics of the generated RNASeq data are tabulated in Table 4.

Tissue	Raw Reads	Total bases (Gb)	High Quality Trimmed reads	Q20	Q30	GC %	Mean Read length (bp)
Liver	76,700,790	11.40	64,910,604	95.81	91.04	41.78	148.65
Stomach	76,541,758	11.92	70,053,760	96.89	92.22	45.64	155.75
Ovary	76,000,416	10.49	65,766,122	98.50	95.49	49.46	138.05
Kidney	73,497,852	9.98	44,222,196	91.42	84.95	38.95	135.75
Brain	69,777,610	10.62	61,312,996	96.14	91.29	41.14	152.15
Muscle	64,226,884	9.98	57,043,626	96.16	91.58	44.65	155.30
Eyes	53,099,692	7.74	41,344,326	94.47	89.43	45.25	145.75
Spleen	52,743,924	7.90	34,745,044	89.73	82.43	36.23	149.85
Skin	50,825,798	7.74	34,709,164	90.33	83.21	36.66	152.30
Gills	40,014,608	6.19	19,215,460	83.89	74.01	29.48	154.60
Dorsal Barbel	29,407,824	4.43	18,470,024	93.04	86.32	44.68	150.65
Gall Bladder	23,284,612	3.51	17,325,114	95.82	90.24	50.74	150.55
Anal Fin	22,444,658	3.38	18,139,824	95.27	89.23	45.93	150.55
Heart	21,755,684	3.21	15,139,424	91.17	84.78	39.86	147.50
Dorsal Fin	21,347,426	3.21	16,314,830	95.41	89.45	46.01	150.55
Pectoral Fin	20,162,134	3.04	15,883,040	94.68	88.55	45.69	150.55
Arborescent organ	20,097,026	3.03	15,232,480	95.23	89.25	46.86	150.55

Table 4. Raw data statistics for RNASeq.

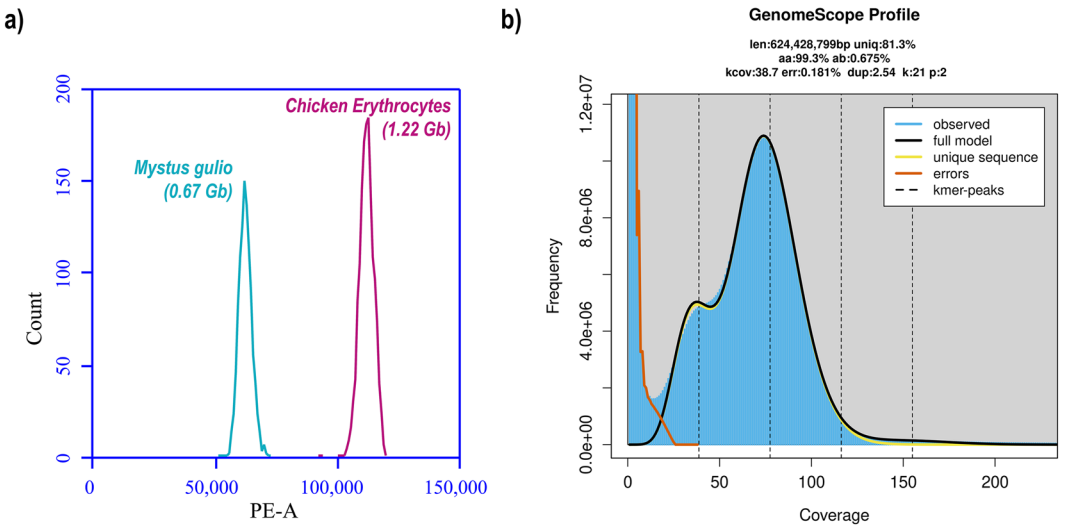

Fig. 3 Genome size estimation profile of *M. gulio*. (**a**) Flowcytometry profile showing *M. gulio* and chicken erythrocytes genomic content peaks. (**b**) Genome size estimated using the k-mer frequency profile generated by using Jellyfish and Genomescope.

Genome size assessment. The genome size was assessed using the blood sample of *M. gulio* using propidium iodide staining[22] on BD Accuri™ C6 Flow cytometer (BD Biosciences, California, USA). The control consisted of chicken erythrocytes from the BD™ DNA QC Particles kit (BD Biosciences, California, USA). The genome size was estimated to be 0.67 Gb based on the histogram data processed using BD Accuri™ C6 Plus software v1.0.23.1 (Fig. 3a). The genome size assessment based on Illumina paired-end DNA short reads sequence data k-mer profiling using k-mer length of 21 bp was performed using Jellyfish v2.2.3, and the counts histograms were later visualized using Genomescope v2.0. The analysis revealed the approximate genome size of 624.43 Mb with heterozygosity of 0.675% (Fig. 3b).

Genome assembly. The HiFi reads extracted from each of the three SMRT cells were subjected to contamination checks from sequencing adaptors and foreign contaminants using the NCBI foreign contamination screen tool[23] to obtain clean HiFi reads. The clean HiFi reads from all three SMRT cells were combined to obtain approximately 57.18 Gb (81x coverage). The contig-level draft assembly was obtained using hifiasm v0.24.0-r702[24] with the clean HiFi reads. The assembly was further examined for the presence of foreign contaminant sequences using the NCBI foreign contamination screen tool[23]. The final contig assembly was 706.31 Mb in length, comprising 283 contigs, with an N50 value of 14.99 Mb. The Hi-C reads were pre-processed using the fastp v1.0.1 tool[25] to remove the adaptor content and low-quality reads. The final scaffolding was performed by utilising the pre-processed

	Contigs	Scaffolds
Number	283	108
Total Length, bp	706,317,132	706,324,109
Longest sequence, bp	32,465,717	40,006,084
N50, bp	14,992,502	22,795,727
L50	16	13
GC, %	39.43%	39.43%
N's per Kb	—	0.03

Table 5. Assembly statistics for *Mystus gulio*.

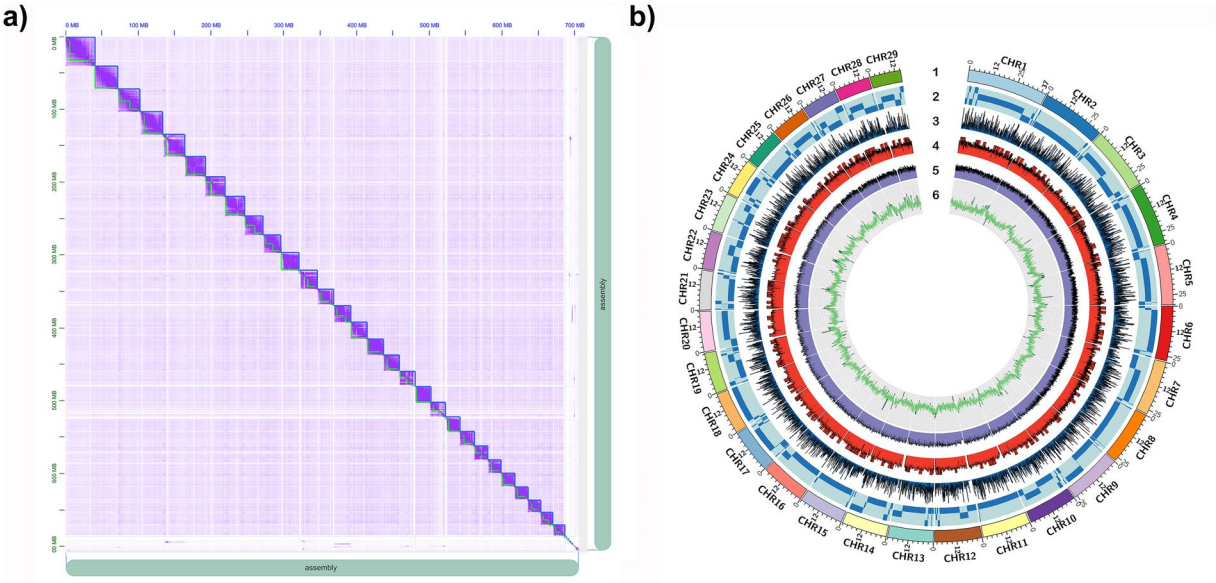

Fig. 4 (**a**) Hi-C contact matrix representing the assembled genome of *M. gulio* in 29 chromosome level scaffolds. (**b**) Circos plot representing various attribute of *M. gulio* genome. From the outermost: Track 1: The 29 chromosomes of *M. gulio* genome. Track 2: Contigs corresponding to the 29 chromosomes represented as tiles. Track 3: The protein encoding genes shown a line diagram. Track 4: Full length isoform sequences supporting the genes of *M. gulio* shown as line diagram. Track 5: Transcripts generated from RNAseq data supporting the genes of *M. gulio* shown as line diagram. Track 6: GC content of *M. gulio* genome shown as a line diagram plotted with 50 kb sliding window. The Gc value between 35 to 45 are shown in green colour.

Hi-C reads and contig assembly by employing the Arima Genomics' mapping pipeline, which includes mapping the individual Hi-C reads to the contig assembly using bwa-mem2 v2.2.1[26] to generate sam files followed by converting the sam files to bam files using samtools v1.21-24[27]. The *filter_five_end.pl* script provided by the Arima Genomics' mapping pipeline was used to remove over-represented 3 bp molecular barcode UMI sequences and 2 dark bases from the 5′ end of reads, followed by using the *two_read_bam_combiner.pl* script to combine the forward and reverse bam files from each of the Hi-C libraries. The '*AddOrReplaceReadGroups, MarkDuplicates.* and *MergeSamFiles*' modules of the Picard tools v3.4.0 (https://broadinstitute.github.io/picard/) were then used to first add read group identifiers to combined bam files followed by marking of duplicated reads and then merging the bam files originated from the two different Hi-C libraries. The final bam files and the contig files were used as inputs for the YaHS v1.1 tool[28] for scaffolding.along with the '*-e GATC,GANTC,CTNAG,TTAA*' option. The scaffold representing the complete mitochondrial genome was identified by blast analysis of the known mitochondrial genome of *M. gulio* (NCBI Accession: NC_069221) against the scaffold assembly. The final scaffold assembly represented the nuclear genome of 706.32 Mb in length, comprising 108 scaffolds, with an N50 value of 22.79 Mb (Table 5) and 1 scaffold representing the mitochondrial genome 16,523 bp in length. *M. gulio* has been reported to contain 29 chromosome pairs[10], and the top 29 scaffolds covered 96.57% of the assembled genome. This indicates a highly contiguous assembly for *M. gulio* (Fig. 4). The chromosomes were further scanned for the presence of telomeric repeat sequence (AACCCT) using the tidk v0.2.0 tool[29]. Telomeric repeats were observed at the ends of the 29 chromosome level scaffold (Fig. 5). Similar assembly strategy combining of PacBio HiFi reads and Hi-C reads data lead to generation of chromosome scale assemblies with telomere ends in recently published brackishwater fish species[30,31].

Repeat masking. The repeat profile of the assembled *M. gulio* genome was established by *de novo* repeat prediction using Repeat Modeler v2.0.5[32]. Repeat Modeller utilises TRF, RECON, and LTR Retriever tools to identify

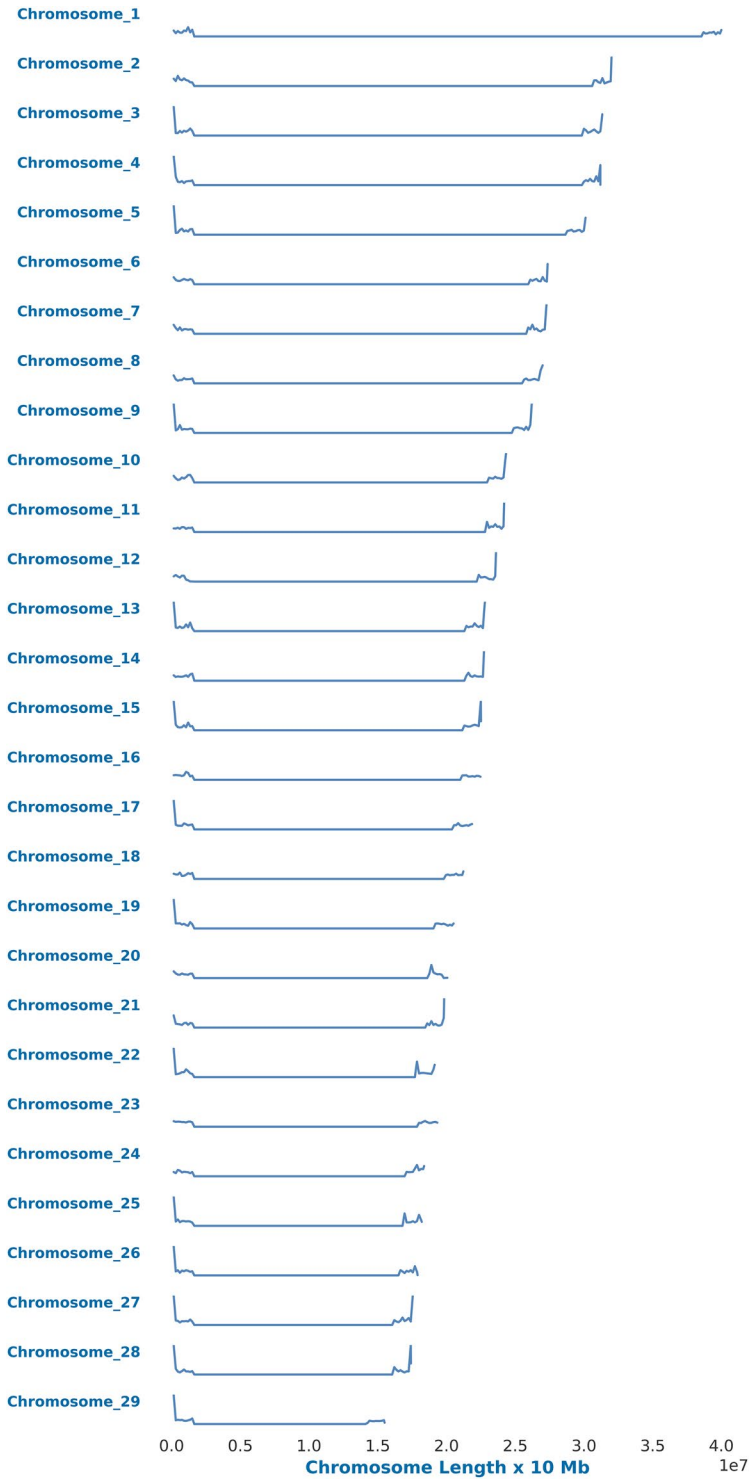

Fig. 5 The graph representing the telomeric ends identified on the assembled chromosome level scaffolds of *M. gulio* genome.

various types of repeat elements. The *de novo* repeat library generated by Repeat Modeler identified 2,174 repeat families and was utilised for masking the repeats in the assembled genome using the Repeatmasker v4.1.5 tool[33]. The total repeat elements covered 32.66% of the assembled genome sequence. The majority of the repeat elements were labelled as unclassified (13.11%), followed by Retroelements (6.64%), Simple repeats (6.53%), and DNA transposons (5.51%). (Table 6). The repeat masked genome, excluding the low complexity and simple sequence repeats, was used for gene prediction.

Sequences	108		
Total length	706,324,109		
GC level	39.43%		
Bases masked	230,697,150 (32.66%)		
Repeat type	**Number of elements**	**Length occupied, bp**	**Percentage of sequence, %**
Retroelements	**145,082**	**46,919,562**	**6.64%**
SINEs:	**28**	**8,155**	**0.00%**
LINEs:	**83,644**	**24,300,474**	**3.44%**
L2/CR1/Rex	57,409	16,438,538	2.33%
R1/LOA/Jockey	1,419	357,035	0.05%
R2/R4/NeSL	46	13,183	0.00%
RTE/Bov-B	12,066	2,280,520	0.32%
L1/CIN4	6,157	2,367,515	0.34%
LTR elements:	**61,410**	**22,610,933**	**3.20%**
BEL/Pao	153	134,394	0.02%
Gypsy/DIRS1	44,041	13,279,668	1.88%
Retroviral	459	322,816	0.05%
DNA transposons	**217,486**	**38,923,025**	**5.51%**
hobo-Activator	15,131	2,276,421	0.32%
Tc1-IS630-Pogo	175,691	31,200,051	4.42%
PiggyBac	153	81,742	0.01%
Tourist/Harbinger	10,843	1,443,522	0.20%
Unclassified:	**596,197**	**92,627,897**	**13.11%**
Total interspersed repeats:		**178,470,484**	**25.27%**
Small RNA:	**22,376**	**2,746,441**	**0.39%**
Satellites:	**1**	**214**	**0.00%**
Simple repeats:	**679,282**	**46,132,947**	**6.53%**
Low complexity:	**49,692**	**3,347,064**	**0.47%**

Table 6. Repeat profile of *Mystus gulio* genome assembly.

Gene prediction and annotation. The gene prediction pipeline included evidence from *ab initio* predictions, RNAseq data and Iso-seq data as described earlier[34,35] with minor changes. The *ab initio* predictions was performed on the repeat masked genome of *M. gulio* as input for Augustus v3.4.0 tool[36] RNA seq reads from 17 different tissues of *M. gulio* were aligned to the assembled genome using Hisat v2.2.1[37] and subsequently assembled to transcripts using Stringtie v2.2.1[38]. The transcripts were further processed with TransDecoder v5.7.0 to identify candidate coding regions evidences for downstream processing. The PacBio long read Iso-seq data processed upto flnc transcripts were aligned to the assembled genome using GMAP v2021-12-17[39] to generate long read hints. Finally Evidence Modeler v2.0[40] was used to combine the evidences in a weighted manner with *ab initio*, *ab initio* with Iso-seq hints, RNAseq hints and Iso-seq hints having weights 3, 9, 4, and 5 respectively. This approach resulted in a consensus gene model containing 23,339 protein encoding genes (Table 7). The homology-based annotation of the predicted proteins was performed using the blastx tool[41] against the Actinopterygii (txid7898) dataset from NCBI's non-redundant database. The protein domain and functional motifs-based annotation was performed using the InterProScan module of the OmicsBox tool[42]. Orthology-based annotation was performed using the EggNOG mapper module[43] of the OmicsBox tool[42]. Finally, the annotations were merged to obtain functional annotations based on Gene Ontology (Table 8). The sequences were mapped to the KEGG pathway database[44] to identify the pathways. A total of 9,161 protein-encoding genes were identified to be part of 338 pathways (List of pathhways, - https://doi.org/10.6084/m9.figshare.30070591.v3).

Isoform classification. The flnc transcripts were aligned to the reference genome using the pbmm2 v1.17.0 (https://github.com/PacificBiosciences/pbmm2), a PacBio-optimized wrapper around minimap2[45], with *ISOSEQ* and *sort* flags enabled. The mapped bam file was then processed using *IsoSeq* v4.3.0[46] which collapses redundant isoforms with biologically significant differences in transcript architecture preserved. The *--do-not-collapse-extra-5exons* option was used to retain transcript diversity in the 5' UTRs, which may reflect alternative promoter usage. The transcript classification was performed using the Pigeon: Iso-Seq Transcript Classification Toolkit v1.1.0 (https://isoseq.how/classification/pigeon). First, the *prepare* module was implemented to sort and index genomic features and transcripts for the classification step, ensuring chromosomal contiguity and generating the required index files. Next the *classify* module, which compares novel transcript structures to known reference annotations, was used to produce isoform classification, splice junctions' characterisations, and compiled summary statistics. The gtf file generated in the gene prediction step was used as reference annotation file in this step. To increase the accuracy of isoform quantification and enhance the biological relevance of classification, the previously obtained flnc counts were integrated. This classification was then further filtered to obtain a filtered classification and junction files. The

Property		Value
All	Count	23,339
	Transcripts Per Gene	1
	Average Length	13,468
	Median Length	7,537
	Total Length	314,334,487
	Average Coding Length	1,479
	Median Coding Length	1,083
	Total Coding Length	34,518,297
	Total Exons	207,774
Complete	Count	18,340
	Average Length	11,132.72
	Median Length	6,751
	Total Length	204,174,069
	Average Coding Length	1,476
	Median Coding Length	1,128
	Total Coding Length	27,069,332
	Total Exons	161,156
Incomplete	Count	4,999
	Average Length	22,332
	Median Length	14,631
	Total Length	110,160,418
	Average Coding Length	1,490
	Median Coding Length	848
	Total Coding Length	7,448,965
	Total Exons	46,618

Table 7. Protein encoding genes prediction statistics.

Attributes	No. of Transcripts	Percentage
Predicted Genes	23,339	100.00
Genes with InterProScan Hits	19,374	83.01
Genes with Blast Hits	20,659	88.52
Genes with EggNOG annotation	12,444	53.32
Genes with GO Annotation	18,717	80.19
Genes with GO Mapping	15,509	66.45
Genes with Enzyme code	7,171	30.72
Annotated genes (overall)	21,206	90.86

Table 8. Annotation statistics of predicted protein encoding genes.

isoforms were categorized into several types such as Antisense, Full splice match, Genic genomic, Genic intron, Incomplete splice match, Intergenic, Novel in catalog, Novel not in catalog and others (Fig. 6).

Alternate splicing analysis. SUPPA2 v2.3[47], a tool for quantifying transcript isoforms, was employed to analyse alternative splicing (AS) events[48]. AS events were obtained from the filtered and collapsed isoform annotation formatted in GFF3. The major classes of splicing events consisted Skipped Exon (SE), Alternative 5'/3' Splice Site (A3 and A5 respectively), Mutually Exclusive Exons (MX), Retained Intron (RI), and Alternative First/Last Exon (AF and AL). This analysis provided information on transcriptome complexity through splicing events to different tissues. (Fig. 7)

Ethics statement. The animal study was reviewed and approved by the Institute Animal Ethics Committee (CIBA/IAEC/2023-1 dated 15.02.2023) of ICAR—Central Institute of Brackishwater Aquaculture, Chennai, India.

Data Records

All the raw datasets were deposited under the NCBIs Bioproject accession PRJNA1293097[49] and linked to SRA Study under accession SRP607064[50]. The raw datasets generated using Illumina short read are deposited in NCBI's SRA database with the accession numbers SRR34890038-045, and SRR34890049-057 for RNASeq data, SRR34890046 and SRR34890047 for Hi-C Data and SRR34890045 for WGS data. The HiFi genomic read dataset generated using PacBio Long read sequencing are deposited in NCBI's SRA database with the accession numbers

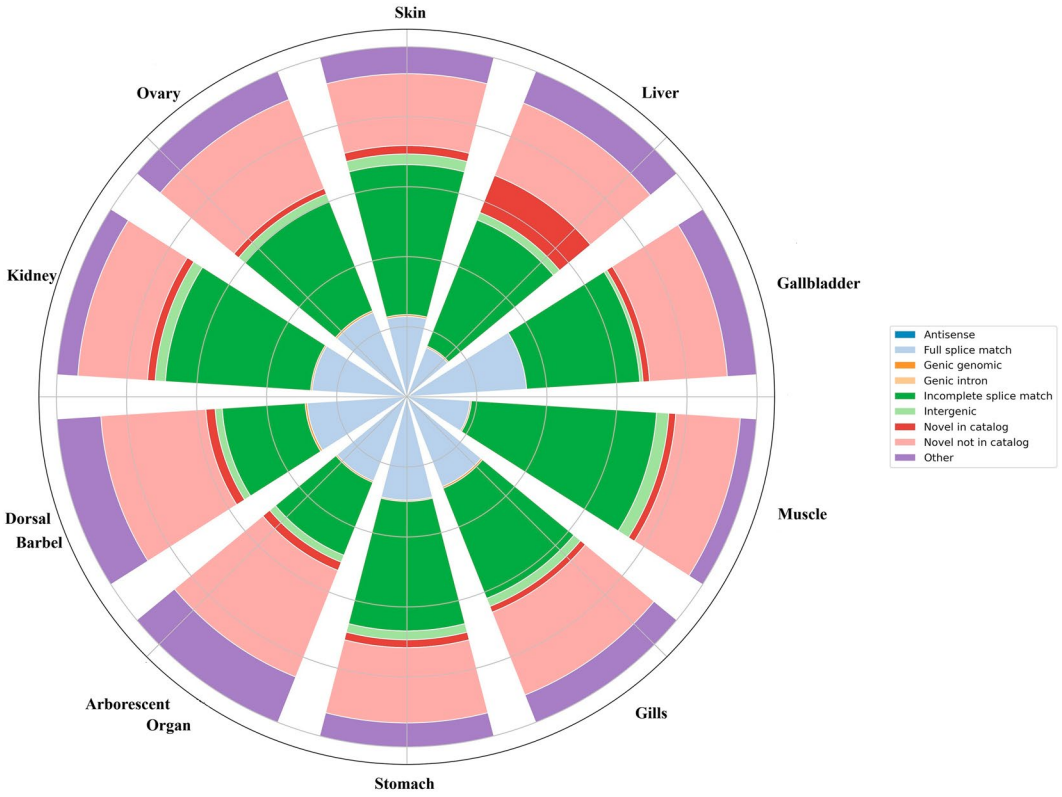

Fig. 6 An illustration of the classification of isoforms in *M. gulio* across various tissues in a stacked chart. The relative distribution of transcript isoforms is represented by stacked coloured bands, with each sector indicating a different tissue type (skin, liver, gallbladder, muscle, gills, stomach, arborescent organ, dorsal barbel, kidney and ovary). Tissue-specific transcriptome patterns are revealed by this visualisation, which emphasises the contribution of several isoform types within each tissue.

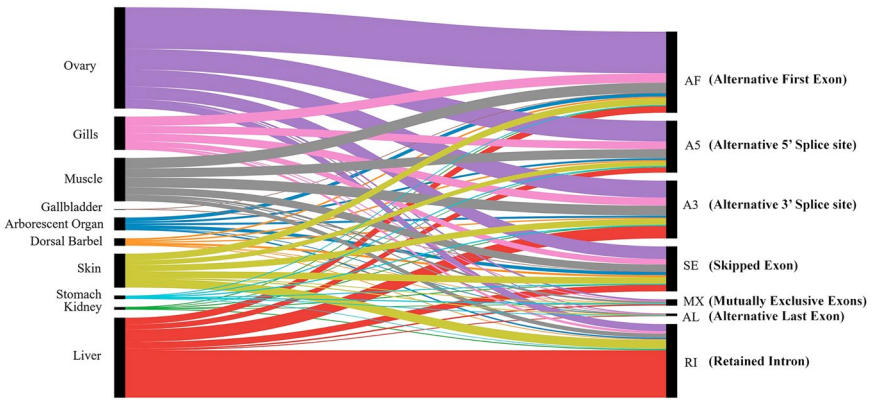

Fig. 7 Linkage map showing the alternative splicing events linked to different tissues of *M. gulio*. The left panel shows the tissue types and the right panel indicates the alternative splicing events. The width of the ribbon shows the absolute count abundance between each tissue and alternative splicing events.

SRR34946063, SRR34946067 and SRR34946068. The FLNC reads generated with PacBio Long read sequencing for Iso-Sequencing are deposited in NCBI's SRA database with the accession numbers SRR34946056-062, and SRR34946064-066. The assembled genome was submitted under the genome category of NCBI with accession number GCA_054643115.1[51]. The full-length transcript files have been uploaded to NCBI's TSA database under the NCBIs Bioproject accession PRJNA1293097[49]. The contig assembly, genome assembly, predicted genes, annotations, Iso-Seq Full length transcripts, and Splicing variants output of SUPPA tool are uploaded to the Figshare repository[52] https://doi.org/10.6084/m9.figshare.30070591.v3.

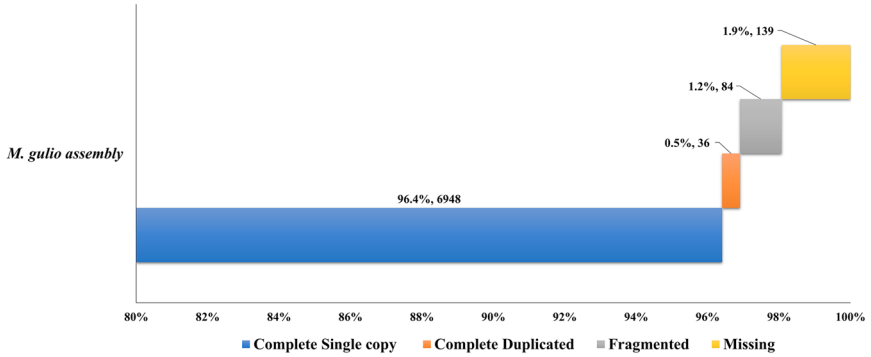

Fig. 8 Graph depicting the results of Genome completeness based on BUSCO analysis of *M. gulio* genome against Actinopterygii_odb12 orthologous database.

S. No.	Data	Alignment rate (%)
1	DNA - Illumina short reads	99.00
2	RNA - Arborescent Organ	96.36
3	RNA - Anal fin	96.03
4	RNA - Brain	94.35
5	RNA - Dorsal Barbel	95.96
6	RNA - Dorsal Fin	96.97
7	RNA - Eyes	92.49
8	RNA - Gall Bladder	98.37
9	RNA - Gills	80.60
10	RNA - Heart	83.53
11	RNA - Kidney	84.48
12	RNA - Liver	91.54
13	RNA - Muscle	93.72
14	RNA - Ovary	99.98
15	RNA - Pectoral Fin	95.18
16	RNA - Skin	93.81
17	RNA - Spleen	96.40
18	RNA - Stomach	98.83

Table 9. Read alignment of Illumina short read DNA and RNA datasets generated in this study against the *M. gulio* genome.

Technical Validation

The genome assembly contiguity and quality was assessed based on the following parameters. (1) the N50 length and L50 values, (2) the presence of full-length mitochondrial DNA in a single scaffold, (3) percentage of the assembled genome in the top longest scaffolds equivalent to the chromosome number of *M. gulio*, (4) the presence of telomeric repeats in the assembled scaffolds, (5) the genome completeness assessed BUSCO v5.7.0 tool[53], (6) the quality value (QV) and error rate in the genome calculated by using Merqury v1.3 tool[54] and (7) raw read representation statistics 8) synteny with closely related species.

The highly contiguous assembly at contig level consisted of 283 contigs with N50 length of 14.99 Mb and L50 value of 16. The scaffold level assembly consisted of 108 scaffolds with an N50 length of 22.79 Mb and L50 value of 13, indicating a reference level assembly considering a chromosome number of 29 for *M. gulio*. The full-length mitochondrial genome was observed in a single scaffold of 16,523 bp. The assembled genome indicated a chromosome-level assembly with about 96.56% of the assembly length encompassed by the first 29 scaffolds. Telomere identification using tidk v0.2.0 tool[29] revealed the chromosomes contained telomeric ends (Fig. 5). The completeness assessment using BUSCO v5.7.0[53] tool against the actinopterygii_odb12 (2025-04-11) ortholog database revealed a highly complete genome of 98.1%, consisting of 96.9% (6984) of complete orthologs, which includes 96.4% (6948) single copy orthologs and 0.5% (36) duplicate orthologs, and 1.2% (84) of fragmented orthologs and only 1.9% (139) of missing orthologs (Fig. 8). The quality value and error rate evaluated by Merqury v1.3 tool was 59.49 and 1.12e-06, respectively. This implies the possibility of only 1 error base in 8.9×10^5 bases, indicating in a highly accurate genome assembly. The assembly was further validated by the alignment percentage obtained from aligning RNA-seq reads and DNA short reads onto the genome (Table 9). Synteny of *M. gulio* against *Hemibagrus wyckioides* (NCBI accession No: GCF_019097595) and *Tachysurus*

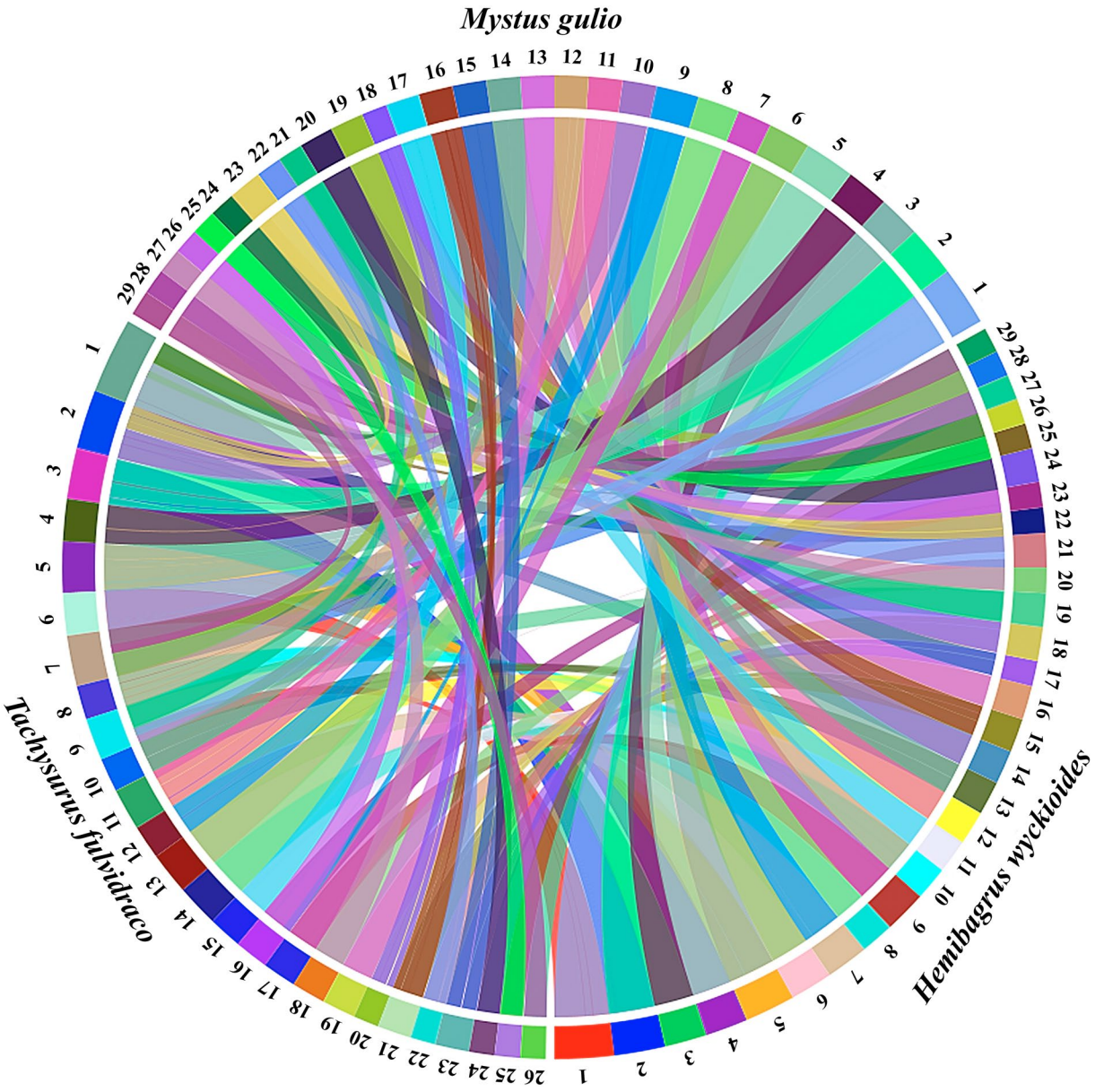

Fig. 9 Three-way synteny plot between the chromosomes of *M. gulio* (29 chromosomes), *Hemibagrus wyckioides* (29 chromosomes) and *Tachysurus fulvidraco* (26 chromosomes).

fulvidraco (NCBI accession No: GCF_022655615) was performed using Symap v5.3.5[55] (Fig. 9). Synteny between *M. gulio* (29 chromosomes) and *H. wyckioides* (29 chromosomes) showed chromosome to chromosome mapping. The synteny of *M. gulio* and *H. wyckioides* against *T. fulvidraco* showed more than one chromosome of *M. gulio* (29 chromosomes) and *H. wyckioides* (29 chromosomes) mapping to single chromosomes of *T. fulvidraco* (26 chromosomes) owing to the difference in chromosome numbers among the three species. The coverage statistics showed 98% similarity between *M. gulio* and *H. wyckioides*, and 97% similarity between *M. gulio* and *T. fulvidraco*. A summary of all the key attributes of the assembly is tabulated in Table 10.

Data availability

The raw datasets generated using Illumina short read sequencing was deposited under the Sequence read archives with accession numbers, SRR34890038-057 and sequences generated using PacBio Long read sequencing were deposited under the accession numbers SRR34946056-068. The assembled genome was submitted under the genome category of NCBI with accession number JBPSLC000000000. The contig assembly, genome assembly, predicted genes and annotations are uploaded to the Figshare repository https://doi.org/10.6084/m9.figshare.30070591.v3.

S. No.	Attributes	Values
1	Scaffold assembly size (bp)	706,324,109
2	Total number of Scaffolds (Nos)	108
3	Scaffold N50 (bp)	22,795,727
4	Mitochondrial genome size (bp)	16,523
5	Total repeat content (%)	32.66
6	Protein encoding genes predicted (Nos)	23,339
7	Total annotated genes [Nos, (%)]	21,206 (90.86%)
8	BUSCO completeness (%)	98.1
9	QV	59.49
10	Error Rate	1.12e-06
11	Alignment rate of Illumina short read DNA data (%)	99.00
12	Average alignment rate of RNASeq data (%)	93.45

Table 10. Summary of the key assembly attributes of *Mystus gulio*.

Code availability

All data processing commands and pipelines were carried out per the instructions and guidelines provided by the relevant bioinformatics software. No custom scripts or code were utilised in this study.

Received: 22 September 2025; Accepted: 26 January 2026;
Published online: 04 February 2026

References

1. Talwar, P. K. & Jhingran, A. G. *Inland Fishes of India and Adjacent Countries*. **vol. 2** (CRC press, 1991).
2. Roos, N., Wahab, M. A., Chamnan, C. & Thilsted, S. H. The Role of Fish in Food-Based Strategies to Combat Vitamin A and Mineral Deficiencies in Developing Countries1. *J Nutr* **137**, 1106–1109 (2007).
3. Bogard, J. R. *et al*. Inclusion of small indigenous fish improves nutritional quality during the first 1000 days. *Food Nutr Bull* **36**, 276–289 (2015).
4. Zaman, S. & Ahammed, T. Evaluation of nutrient composition of selected small indigenous fish species in Bangladesh and potential contribution to nutrient intakes. *Journal of Food Composition and Analysis* **136**, 106814 (2024).
5. Patra, M. K., Acharjee, S. K. & Chakraborty, S. K. Conservation categories of siluroid fishes in North-East Sundarbans, India. *Biodivers Conserv* **14**, 1863–1876 (2005).
6. Mijkherjee, M., Praharaj, A. & Das, S. Conservation of endangered fish stocks through artificial propagation and larval rearing technique in West Bengal, India. *Aquaculture Asia* **7**, 8–11 (2002).
7. Alam, M. J., Begum, M., Islam, M. A. & Pal, H. K. Induced spawning and fry production of nona tengra, *Mystus gulio* (Hamilton.). *Progress. Agric* **17**, 235–238 (2006).
8. Kumar, P., Biswas, G., Ghoshal, T. K., Kailasam, M. & Vijayan, K. K. Embryonic and larval developments of brackish water catfish, *Mystus gulio* (Hamilton and Buchanan, 1822) induced with human chorionic gonadotropin and consequent larval rearing. *Aquac Res* **49**, 2466–2476 (2018).
9. Kumar, P. *et al*. Current knowledge on the biology, captive breeding and aquaculture of the brackishwater catfish, *Mystus gulio* (Hamilton, 1822): A review. *Aquaculture* **499**, 243–250 (2019).
10. Manna, G. K. & Khuda-Bukhsh, A. R. Karyomorphological Studies in Three Species of Teleostean Fishes. *Cytologia (Tokyo)* **43**, 69–73 (1978).
11. Shao, F. *et al*. Chromosome-level genome assembly of the Asian red-tail catfish (*Hemibagrus wyckioides*). *Front Genet* **12**, 747684 (2021).
12. Gong, G. *et al*. Chromosomal-level assembly of yellow catfish genome using third-generation DNA sequencing and Hi-C analysis. *Gigascience* **7**, giy120 (2018).
13. Zhu, C. *et al*. Insights into chromosomal evolution and sex determination of *Pseudobagrus ussuriensis* (Bagridae, Siluriformes) based on a chromosome-level genome. *DNA Research* **29**, dsac028 (2022).
14. Gong, G. *et al*. A chromosome-level genome assembly of the darkbarbel catfish *Pelteobagrus vachelli*. *Sci Data* **10**, 598 (2023).
15. Liu, H. *et al*. A chromosome-level genome assembly of the male darkbarbel catfish (*Pelteobagrus vachelli*) using PacBio HiFi and Hi-C data. *Sci Data* **12**, 351 (2025).
16. He, W.-P. *et al*. Chromosome-level genome assembly of the Chinese longsnout catfish *Leiocassis longirostris*. *Zool Res* **42**, 417 (2021).
17. 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).
18. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol* **35**, 1547–1549 (2018).
19. Rambaut, A. FigTree. *Tree Figure Drawing Tool*. http://tree.bio.ed.ac.uk/software/figtree/ (2009).
20. Katneni, V. K. *et al*. Novel Isoform Sequencing Based Full-Length Transcriptome Resource for Indian White Shrimp, *Penaeus indicus*. *Front Mar Sci* **7**, 1–4 (2020).
21. Shekhar, M. S. *et al*. Genome assembly, Full-length transcriptome, and isoform diversity of Red Snapper, *Lutjanus argentimaculatus*. *Sci Data* **11**, 796 (2024).
22. 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).
23. Astashyn, A. *et al*. Rapid and sensitive detection of genome contamination at scale with FCS-GX. *Genome Biol* **25** (2024).
24. 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).
25. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
26. Vasimuddin, M., Misra, S., Li, H. & Aluru, S. Efficient architecture-aware acceleration of BWA-MEM for multicore systems. in *2019 IEEE international parallel and distributed processing symposium (IPDPS)* 314–324 (2019).
27. Li, H. *et al*. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
28. Zhou, C., McCarthy, S. A. & Durbin, R. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics* **39** (2023).

29. Brown, M., la Rosa, P. M. & Mark, B. A Telomere Identification Toolkit. *Zenodo* https://doi.org/10.5281/zenodo.10091385 (2023).
30. Katneni, V. K. *et al*. Genome assembly at chromosome scale with telomere ends for Pearlspot, *Etroplus suratensis*. *Sci Data* **11**, 1226 (2024).
31. Shekhar, M. S. *et al*. Chromosome-level genome assembly with telomeric repeats at scaffold ends for *Rhabdosargus sarba*. *Sci Data* **12**, 1202 (2025).
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. Chen, N. Using Repeat Masker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* **5**, 4–10 (2004).
34. Katneni, V. K. *et al*. A Superior Contiguous Whole Genome Assembly for Shrimp (*Penaeus indicus*). *Front Mar Sci* **8** (2022).
35. 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).
36. Hoff, K. J. & Stanke, M. Predicting Genes in Single Genomes with AUGUSTUS. *Curr Protoc Bioinformatics* **65**, 1–54 (2019).
37. 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).
38. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290–295 (2015).
39. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
40. 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).
41. 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).
42. Omicsbox. OmicsBox-Bioinformatics made easy (Version 3.0.25) (2019).
43. Huerta-Cepas, J. *et al*. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* **47**, D309–D314 (2019).
44. 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).
45. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
46. Topfer, A. & others. *IsoSeq v3: scalable de novo isoform discovery* (2018).
47. Trincado, J. L. *et al*. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol* **19**, 1–11 (2018).
48. Alamancos, G. P., Pagès, A., Trincado, J. L., Bellora, N. & Eyras, E. Leveraging transcript quantification for fast computation of alternative splicing profiles. *Rna* **21**, 1521–1531 (2015).
49. *NCBI BioProject*. http://identifiers.org/bioproject:PRJNA1293097 (2025).
50. *NCBI Sequence Read Archive*. http://identifiers.org/insdc.sra:SRP607064 (2025).
51. *NCBI Genbank* http://identifiers.org/insdc.gca:GCA_054643115.1 (2026).
52. Sudheesh, K. *et al*. Chromosome level genome and full length transcriptome of long whiskers catfish, *Mystus gulio* (Hamilton, 1822), *Figshare*, https://doi.org/10.6084/m9.figshare.30070591.v3 (2025).
53. 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).
54. 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).
55. 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 Brackish water Aquaculture' funded by Department of Biotechnology, Government of India (BT/PR47080/AAQ/3/1050/2022).

Author contributions

M.S.S., V.K.K. and A.K.J. designed the study. M.S.S., V.K.K. and A.K.J. led the research. S.K.P, V.K.K., A.S. and M.S.S. wrote the draft manuscript. S.K.P., K.K. and A.S., 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 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 M.S.S.