


A high-quality chromosome-level genome assembly of rohu carp, *Labeo rohita*, and its utilization in SNP-based exploration of gene flow and sex determination

Mark A. Arick II ^{1,*} Corrinne E. Grover,² Chuan-Yu Hsu,¹ Zenaida Magbanua,¹ Olga Pechanova,¹ Emma R. Miller,² Adam Thrash,¹ Ramey C. Youngblood,¹ Lauren Ezzell,¹ Md Samsul Alam,³ John A.H. Benzie,⁴ Matthew G. Hamilton,⁴ Attila Karsi,⁵ Mark L. Lawrence,⁶ Daniel G. Peterson¹

¹Institute for Genomics, Biocomputing & Biotechnology, Mississippi State University, Mississippi State, MS 39762, USA

²Ecology, Evolution, and Organismal Biology Department, Iowa State University, Ames, IA 50010, USA

³Department of Fisheries Biology and Genetics, Bangladesh Agricultural University, Mymensingh 2202, Bangladesh

⁴WorldFish, Jalan Batu Maung, 11960 Bayan Lepas, Penang, Malaysia

⁵Department of Comparative Biomedical Sciences, College of Veterinary Medicine, Mississippi State University, Mississippi State, MS 39762, USA

⁶Global Center for Aquatic Health and Food Security, Mississippi State University, Mississippi State, MS 39762, USA

*Corresponding author: Institute for Genomics, Biocomputing & Biotechnology, Mississippi State University, Mississippi State, MS 39762, USA. Email: maa146@igbb.msstate.edu

Abstract

Labeo rohita (rohu) is a carp important to aquaculture in South Asia, with a production volume close to Atlantic salmon. While genetic improvements to rohu are ongoing, the genomic methods commonly used in other aquaculture improvement programs have historically been precluded in rohu, partially due to the lack of a high-quality reference genome. Here we present a high-quality de novo genome produced using a combination of next-generation sequencing technologies, resulting in a 946 Mb genome consisting of 25 chromosomes and 2,844 unplaced scaffolds. Notably, while approximately half the size of the existing genome sequence, our genome represents 97.9% of the genome size newly estimated here using flow cytometry. Sequencing from 120 individuals was used in conjunction with this genome to predict the population structure, diversity, and divergence in three major rivers (Jamuna, Padma, and Halda), in addition to infer a likely sex determination mechanism in rohu. These results demonstrate the utility of the new rohu genome in modernizing some aspects of rohu genetic improvement programs.

Keywords: rohu, genome, ddRAD-seq, aquaculture, rui, sex determination

Introduction

Labeo rohita (rohu; rui), a carp naturally found in the Indo-Gangetic and surrounding river systems (Das et al. 2020), is an important aquaculture fish in many areas of South Asia (FAO 2020). The annual aquaculture production of *L. rohita* in Bangladesh was 386.3 thousand tonnes in the 2019–2020 fiscal year, the second-highest among all aquaculture species in the country (DoF 2020). Annual aquaculture production of the species is approximately 2.0 million metric tons (Mt) globally, a volume comparable with *Salmo salar* (Atlantic salmon; 2.4 Mt); however, study and understanding of *L. rohita* genomics is not commensurate with its global significance (Rasal and Sundaray 2020). Although there is increasing interest in applying next-generation sequencing (NGS) and other high-throughput methods to *L. rohita* (Robinson et al. 2014; Rasal et al. 2017; Hamilton et al. 2019; Rasal et al. 2020; Sahoo et al. 2021), to date, most studies have been conducted in the absence of a genome sequence. Recently, a draft genome was published for *L. rohita* (Das et al. 2020) to provide a unifying resource for

NGS analysis; however, the quality of the genome limits the development of a robust genomic framework for the species.

Genetically improved *L. rohita* seed is increasingly available to farmers, from both mass-selection (e.g. “Subarna Rohu” in Bangladesh and “Ayeyarwady Hatchery” in Myanmar) (Hamilton 2019; SZA 2021) and family-based (i.e. pedigree-based) improvement programs (e.g. “Jayanti” in India and “WorldFish Genetically Improved Rohu” in Bangladesh) (Das Mahapatra et al. 2007; Rasal et al. 2017; Hamilton et al. 2019; Hamilton et al. 2022). However, genomic methods routinely applied in other aquaculture species (e.g. parentage assignment and genomic selection) have yet to be routinely applied in *L. rohita* genetic improvement programs (Sahoo et al. 2017; Rasal and Sundaray 2020), primarily due to a historical focus on improving growth rate (directly assessable at low cost on selection candidates), limited financial resources, and the absence of a genome sequence. As existing family-based programs expand to include additional traits (e.g. carcass traits, feed conversion ratio, tolerance to extreme environments, and disease resistance) (Rasal and Sundaray 2020), the advantages afforded by

Received: September 08, 2022. Accepted: December 16, 2022

© The Author(s) 2023. Published by Oxford University Press on behalf of the Genetics Society of America.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

improved genomic resources in *L. rohita* will become increasingly compelling.

The mechanism of sex determination (SD) in *L. rohita* is a lingering question with applications to aquaculture, as understanding SD mechanisms in other species has been used to prevent precocious maturation, exploit sexual dimorphism in growth rate, improve carcass quality, and protect both environmental values and intellectual property (Budd *et al.* 2015). Despite its relevance to aquaculture and genetic improvement, SD in *L. rohita* has been understudied (Sahoo *et al.* 2021) both due to the high diversity of teleost SD mechanisms (Heule *et al.* 2014) and the lack of high-quality genomic resources (Sahu *et al.* 2013).

Here we present a new de novo high-quality genome for *L. rohita* that improves sequence contiguity and reduces duplication. We use this reference to assess diversity among populations of *L. rohita* from three different rivers and to preliminarily describe the gametic system of SD in *L. rohita*, demonstrating the utility of this improved sequence to increase understanding and facilitate aquacultural production and genetic improvement.

Materials and methods

Sample collection, DNA extraction, and sequencing

Blood samples were collected from five male *Labeo rohita* (henceforth referred to as Rohu-1 through Rohu-5) from a fish farm located in the District of Rangpur, Bangladesh. The fish were handled as per guidelines of the Ethics Standard Review Committee of Bangladesh Agricultural University (BAU) involving fish and animals (approval no. BAURES/ESRC/2019/Fish/01). Each fish was euthanized using clove oil, dissected, and blood was collected from the heart using a syringe. Each blood sample was placed in an ethylenediaminetetraacetic acid containing vial, and vials were shipped in an insulated container to Mississippi State University for DNA extraction.

High-molecular-weight (HMW) genomic DNA for whole genome sequencing was extracted from 150 μ l of blood from Rohu-1 using CTAB lysis buffer followed by the phenol/chloroform purification procedure (Doyle and Doyle 1987). The concentration and purity of extracted genomic DNA samples were measured by a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA). The quality of genomic DNA was validated by electrophoresis on a 0.8% w/v agarose gel.

The genomic DNA from Rohu-1 was used to prepare 10 Oxford Nanopore R9.4 MinION flow cells. For each flow cell, 2 to 2.5 μ g of genomic DNA and a Nanopore Genomic DNA Ligation Sequencing Kit SQK-LSK 109 (Oxford Nanopore Technologies, Oxford, UK) were used to create a DNA library. For each of the 10 libraries, 700–750 ng of DNA was loaded onto a Nanopore Flow Cell R9.4.1 (Oxford Nanopore Technologies, Oxford, UK) and sequenced on a GridION sequencer (Oxford Nanopore Technologies, Oxford, UK) for 48 h.

Rohu-1 genomic DNA was also sequenced on an Illumina HiSeq X-Ten (2 \times 150 bp). In brief, 2 μ g of Rohu-1 genomic DNA was used with an Illumina TruSeq DNA PCR-free Library Prep Kit (Illumina, San Diego, CA, USA) to create an Illumina sequencing library. The final DNA-Seq library, which had an insert size range of 350–450 bp, was submitted to Novogene (www.en.novogene.com) for two lanes of PE150 on an Illumina HiSeq X-Ten (Illumina, San Diego, CA, USA) sequencer.

A Hi-C library also was prepared using 100 μ l of Rohu-1 blood with the Proximo Hi-C Animal Kit (Phase Genomics, Seattle, WA,

USA). The final Hi-C DNA-Seq library was submitted to Novogene (www.en.novogene.com) for one lane of PE150 Illumina HiSeq X-Ten (Illumina, San Diego, CA, USA) sequencing.

Lastly, Rohu-1 blood cells were embedded in agarose and HMW DNA was isolated according to the Bionano Prep Frozen Blood Protocol (Bionano Genomics, San Diego, CA). The extracted DNA molecules were labeled with the Direct Label and Stain (DLS) DNA Labeling kit (Bionano Genomics, San Diego, CA). Once labeled and stained, the DNA was imaged on the Bionano Saphyr instrument (Bionano Genomics, San Diego, CA).

Genome size estimation

The genome size of *L. rohita* was estimated using two independent methods: flow cytometry and k-mer profiling.

Flow cytometry was performed using erythrocyte nuclei from Rohu-1, Rohu-2, Rohu-3, Rohu-4, and Rohu-5 using trout erythrocyte nuclei (TENS; https://www.biosure.com/tens.html) as a standard (1C = 6.5 pg). For each replicate, nuclei were stabilized in 200 μ l of LB01-propidium iodide (PI) buffer as per (Pellicer and Leitch 2014), and two drops of TENS standard were used per 50 μ l of fish blood. Each sample was measured twice, totaling 10 runs overall. Only measurements with greater than 5,000 nuclei and a coefficient of variation (CV) of less than 3% were retained (Pellicer and Leitch 2014).

For k-mer profiling, Jellyfish [v2.2.10] (Marçais and Kingsford 2011) was used to “digest” the Rohu-1 Illumina paired reads into 50-mers. GenomeScope [v1.0] (Vurture *et al.* 2017) was then used to estimate genome size using the resulting k-mer profile.

Assembly and annotation

Nanopore sequence data was filtered to remove the control lambda-phage and sequences shorter than 1,000 bases using the nanopack tool suite [v1.0.1] (De Coster *et al.* 2018). Trimmomatic [v0.32] (Bolger *et al.* 2014) was used to remove adapters, trim low-quality bases, and filter out reads shorter than 85 bp. The filtered nanopore data were assembled into contigs using wtdbg2 [v2.4] (Ruan and Li 2020). The contigs were polished using two iterations of racon [v1.4.0] (Vaser *et al.* 2017) with minimap2 [v2.17] (Li 2018) mapping the nanopore reads. The contigs were further polished with Illumina paired-end read data using pilon [v1.23] (Walker *et al.* 2014) with bwa [v0.7.10] (Li 2013) mapping the Illumina paired reads. The resulting contigs were scaffolded using Bionano Solve [Solve3.4.1_09262019] using the optical mapping data generated from the Saphyr run. SALSA [v2.3] (Ghurye *et al.* 2019) was used to produce super-scaffolds using the Hi-C library and the Bionano scaffolded sequences. Those scaffolds larger than 10Mb were linked and oriented based on the *Onychostoma macrolepis* genome (Sun *et al.* 2020), the chromosome assembly most similar to *L. rohita* available on NCBI, using RagTag [v1.1.1] (Alonge *et al.* 2022).

RepeatModeler [v2.0.1] (Flynn *et al.* 2020) and RepeatMasker [v4.1.1] (Smit *et al.* 2013) were used to create a species-specific repeat database, and this database was subsequently used by RepeatMasker to mask those repeats in the genome. All available RNA-seq libraries for *L. rohita* (comprising brain, pituitary, gonad, liver, pooled, and whole body tissues for both sexes; Supplementary Table 1) were downloaded from NCBI and mapped to the masked genome using hisat2 [v2.1.0] (Kim *et al.* 2019). These alignments were used in both the mikado [v2.0rc2] (Venturini *et al.* 2018) and braker2 [v2.1.5] (Brůna *et al.* 2021) pipelines. Mikado uses putative transcripts assembled from the RNA-seq alignments generated via stringtie [v2.1.2] (Kovaka *et al.* 2019), cufflinks [v2.2.1] (Trapnell *et al.* 2012), and trinity [v2.11.0] (Grabherr *et al.*

2011) along with the junction site prediction from portcullis [v1.2.2] (Mapleson et al. 2018), the alignments of the putative transcripts with UniprotKB Swiss-Prot [v2021.03] (The UniProt Consortium 2021), and the ORFs from prodigal [v2.6.3] (Hyatt et al. 2010) to select the best representative transcript for each locus. Braker2 uses those RNA-seq alignments and the gene prediction from GeneMark-ES [v4.61] (Borodovsky and Lomsadze 2011) to train a species-specific Augustus [v3.3.3] (Stanke et al. 2006) model. Maker2 [v2.31.10] (Holt and Yandell 2011) predicts genes based on the new Augustus, GeneMark, and SNAP models derived from Braker2 along with the Mikado predicted transcripts as an external *ab-initio* source, modifying the predictions based on the available RNA and protein evidence from the Cyprinidae family in the NCBI RefSeq database. Any predicted genes with an annotation edit distance (AED) above 0.47 were removed from further analysis. The remaining genes were functionally annotated using InterProScan [v5.47-82.0] (Jones et al. 2014) and BLAST+ [v2.9.0] (Camacho et al. 2009) alignments against the UniprotKB Swiss-Prot database. BUSCO [v5.2.2] (Manni et al. 2021) was used to verify the completeness of both the genome and annotations against the actinopterygii_odb10 database. Lastly, genes spanning large gaps or completely contained within another gene on the opposite strand were removed using a custom Perl script (<https://github.com/IGBB/rohu-genome/>).

Comparative genomics

The assembly statistics, length distributions, BUSCO completeness scores, and sequence similarity via dot-plots were compared between the IGBB *L. rohita* genome (reported here) and the *L. rohita* genome reported by Das et al. (2020) (CIFA, Refseq accession GCA_004120215.1), as well as all 12 annotated Cypriniformes genomes from NCBI (Table 1). Assembly statistics were calculated using abyss-fac from ABySS [v2.3.4] (Jackman et al. 2017). Length distributions were calculated using samtools [v1.9] (Danecek et al. 2021) and graphed using R [v4.0.2] (R Core Team 2020) with the tidyverse package (Wickham et al. 2019). Minimap2 [v2.17-r941] and the pafCoordsDotPlotly R script (<https://github.com/tpoorten/dotPlotly>) were used to create dot-plots. For the Cypriniformes data-set, only chromosome level assemblies were included in the dot-plots. The *Danio rerio* (zebrafish) and *Triplophysa tibetana* genomes were also excluded from the dot-plots since few of the alignments passed the default quality filter in pafCoordsDotPlotly. BUSCO with the actinopterygii_odb10 database was used to find the BUSCO scores for each genome. The annotated genes from this new assembly were also compared to all annotated Cypriniformes using OrthoFinder [v2.5.4] (Emms and Kelly 2019).

Table 1. List of Cypriniformes genomes used in comparative analyses.

Organism scientific name	Assembly name	Assembly accession	L	Contig N50	Size	Submission date	Gene count
<i>Anabarrilius grahami</i>	BGI_Agra_1.0	GCA_003731715.1	S	36.06 Kb	991.89 Mb	2018-11-15	23,906
<i>Carassius auratus</i>	ASM336829v1	GCF_003368295.1	C	821.15 Kb	1820.62 Mb	2018-08-09	83,650
<i>Cyprinus carpio</i>	ASM1834038v1	GCF_018340385.1	C	1.56 Mb	1680.12 Mb	2021-05-12	59,559
<i>Danionella translucida</i>	ASM722483v1	GCA_007224835.1	S	133.13 Kb	735.30 Mb	2019-07-22	35,803
<i>Danio rerio</i>	GRCz11	GCF_000002035.6	C	1.42 Mb	1373.45 Mb	2017-05-09	40,031
<i>Onychostoma macrolepis</i>	ASM1243209v1	GCA_012432095.1	C	10.81 Mb	886.57 Mb	2020-04-17	24,754
<i>Pimephales promelas</i>	EPA_FHM_2.0	GCA_016745375.1	S	295.77 Kb	1066.41 Mb	2021-01-24	26,150
<i>Puntigrus tetrazona</i>	ASM1883169v1	GCF_018831695.1	C	1.42 Mb	730.80 Mb	2021-06-10	40,303
<i>Sinocyclocheilus anshuiensis</i>	SAMN03320099.WGS_v1.1	GCF_001515605.1	S	17.27 Kb	1632.70 Mb	2015-12-14	52,005
<i>Sinocyclocheilus grahami</i>	SAMN03320097.WGS_v1.1	GCF_001515645.1	S	29.35 Kb	1750.27 Mb	2015-12-16	55,200
<i>Sinocyclocheilus rhinoceros</i>	SAMN03320098_v1.1	GCF_001515625.1	S	18.76 Kb	1655.77 Mb	2015-12-14	53,875
<i>Triplophysa tibetana</i>	ASM836982v1	GCA_008369825.1	C	2.83 Mb	652.93 Mb	2019-09-12	24,398

The "L" column is an abbreviation of the assembly level: (S)caffold and (C)hromosome.

ddRAD-seq sample collection and library prep

Fin clips were taken from the founders of the WorldFish Rohu Genetic Improvement Program, as described in Hamilton et al. (2019). A custom R script (<https://github.com/IGBB/rohu-genome/>) was used to minimize sampling putatively related founders (Hamilton et al. 2019). In total, fin clips from 64 male and 56 female *L. rohita* were sampled, sourced from the Halda (39), Jamuna (38), and Padma (43) rivers.

Genomic DNA was extracted from the samples using the Qiagen DNeasy Blood & Tissue Mini kit (Qiagen, Valencia, CA, USA). The concentration and purity of extracted genomic DNA samples were evaluated using a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA). The quality of genomic DNA was validated by electrophoresis on a 0.8% w/v agarose gel. The ddRAD-Seq libraries were made using the method described in Magbanua et al. (2022) with minor modifications. Briefly, NsiI and MspI were used to digest the genomic DNA and the adapters (Supplementary Table 2) were ligated into the digested genomic DNA. Polymerase chain reaction was used to attach the i5 and i7 index primers [Nextera XT Index Kit v2 Set A (FC-131-2001) and Set B (FC-131-2002), Illumina, San Diego, CA] to the ligation products to provide unique dual barcodes to each sample while generating the sequencing libraries. The libraries were submitted to Novogene (www.en.novogene.com) for a total of two lanes of PE150 Illumina HiSeq X-Ten (Illumina, San Diego, CA, USA) sequencing.

Single-nucleotide polymorphism discovery and population analyses

Reads were mapped to the *L. rohita* genome using bwa [v0.7.17]. Single-nucleotide polymorphisms (SNPs) were called over two rounds using the Sentieon pipeline (Kendig et al. 2019) [Spack version sentieon-genomics/201808.01-opfuvzr] and following the DNaseq guidelines. Briefly, SNPs were predicted for the ddRAD-seq samples using the DNaseq pipeline for all samples, and these SNPs were used as known sites during base quality score recalibration (BQSR) in the second iteration of the DNaseq pipeline. The final SNP set was filtered via vcftools [Spack version 0.1.14-v5mvhea] (Danecek et al. 2011) to remove sites with insufficient representation (i.e. present in <90% of samples). The filtered SNP set was used with the R packages LEA (Frichot and François 2015) for the population structure analysis and SNPrelate (Zheng et al. 2012) for the principle component analysis. Nucleotide diversity (π) and divergence (π_{xy} , or d_{xy}) were calculated in 10 kb windows using pixy v1.2.5.beta1 (Korunes and Samuk 2021) run via Miniconda3 [Spack version 4.3.30-qdauevb]. Population differentiation (F_{st}) was also calculated in pixy using

10 kb windows. Output from pixy was processed in R [4.1.1] and visualized using ggplot2 (Wickham 2016). Specific parameters and code can be found at <https://github.com/IGBB/rohu-genome>.

Sex-associated fragments

To find regions of the *L. rohita* genome associated with sex, two-sample Monte Carlo tests comparing the high-quality read mappings for male and female samples were run for each fragment between the two digestion sites. The digestion site fragments for the *L. rohita* genome were found using egads (<https://github.com/IGBB/egads>). The high-quality read mappings for each sample were calculated by first filtering high-quality (mapq \geq 30) mappings using samtools [v1.9] (Danecek et al. 2021), and then using the bedtools [v2.28.0] (Quinlan and Hall 2010) coverage to count the number of mappings to each fragment. Given the maximum selected size (613 bp) and the paired read size (300 bp), fragments with less than half of the sequence covered in a sample were removed from further analysis. Fragments with fewer than 50 samples (90% of the smallest sample group) surviving the filter were removed altogether. The fragment read mappings for each sample were normalized based on the total number of high-quality read mappings within a sample. Permutation tests were run on each fragment for 100,000 replicates, and the resulting *P*-values were adjusted using the Benjamini-Hochberg method. Fragments with an adjusted *P*-value less than 0.05 were considered associated with sex. The commands and code used can be found at <https://github.com/IGBB/rohu-genome/y-link>.

Results and discussion

Genome size estimation

The C-value of *Labeo rohita* was previously reported as 1.99 pg (~1.95 Gb) based on Feulgen densitometry (Patel et al. 2009) and 1.427 Gb in the currently available assembly (Das et al. 2020); however, our flow cytometry results based on five individuals and our k-mer-based genome size estimation suggest that the *L. rohita* genome size is 50–65% the size previously reported by Patel et al. (2009) and Das et al. (2020), respectively. Our flow cytometry results indicate a C-value of 0.99 pg (~0.97 Gb) with a standard

deviation of only 0.02 across all measurements (Supplementary Table 3). Moreover, our k-mer-based estimate using GenomeScope (complete results in Supplementary Table 4 and Supplementary Fig. 1) is 0.97 Gb, the same value determined by our flow cytometry analysis. Lastly, our final genome assembly size for *L. rohita* is 0.95 Gb. Notably, the Feulgen densitometry estimate reported in Patel et al. (2009) for a second fish, *Labeo catla* (synonymous with *Catla catla*), was also approximately twice that later reported (Sahoo et al. 2020), perhaps suggesting stochastic differences, including cryptic variation in ploidy and/or differences in measurement techniques (Greilhuber 2005). Figure 1 shows the genome size comparison of all samples mentioned above.

Genome assembly and annotation

Genome assembly was started with (a) a total of 130.5 Gb (138X coverage) of Nanopore data, derived from 44.7 million reads, (b) 261 Gb (276X coverage) of Illumina short reads (870 million 150 bp paired-end reads), and (c) 382 million 150 bp paired reads (114 Gb) from a Hi-C library. The initial de novo assembly was generated using the Nanopore data and polished with the short insert Illumina data, resulting in 4,999 contigs with an N50 of 1.28 Mb. After the Bionano and Hi-C data were incorporated, the total number of sequences dropped to 2,899 and the N50 increased to 29.9 Mb. These sequences were ordered and oriented by RagTag using the *Onychostoma macrolepis* reference to produce a final assembly with 25 chromosome-length scaffolds (deemed Chr01 through Chr25—Supplementary Table 5) and 2,844 unplaced scaffolds, which ranged in size from 1,479 bp to 7.18 Mb. The chromosome scaffolds were composed of one to eight sequences each, with all but three composed of three or fewer sequences. The final assembled genome size was 945.5 Mbp, representing 97.9% of the estimated genome size (see Table 2 for assembly statistics at each step).

RepeatModeler2 predicted 3,851 repeat families. Interestingly, while over three-quarters of the predicted TEs remain uncategorized (due to lack of related representatives), *L. rohita* has a relative abundance of LTR-retrotransposons vs other types of elements (e.g. LINES and Class II elements; 730 vs <100 each), which is in contrast to the model fish (i.e. *D. rerio*), where DNA elements are

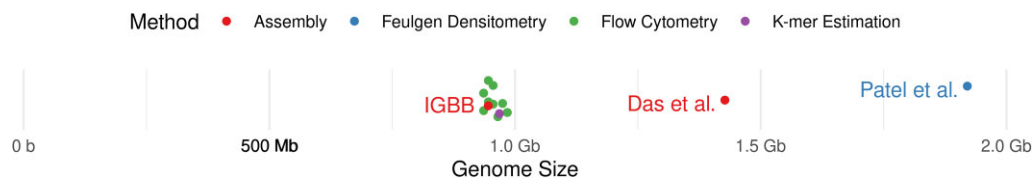


Fig. 1. Genome size estimates among the reported methods.

Table 2. Assembly statistics for each stage of the IGBB *L. rohita* assembly and the CIFA *L. rohita* assembly.

	IGBB				CIFA final
	Nanopore + Illumina	+ Bionano	+ Hi-C	Final	
Number of sequences	4,999	3,709	2,899	2,872	13,623
L50	202	15	14	13	182
Smallest sequence	1,348	1,479	1,479	1,479	—
N75	514,919	11.3 Mb	26.4 Mb	28.8 Mb	774.7 Kb
N50	1.28 Mb	26.5 Mb	29.9 Mb	32.5 Mb	2.01 Mb
N25	2.40 Mb	30.8 Mb	34.3 Mb	36.1 Mb	4.28 Mb
E-size	1.73 Mb	22.0 Mb	26.9 Mb	30.0 Mb	2.91 Mb
Largest sequence	7.83 Mb	37.9 Mb	44.5 Mb	45.3 Mb	15.2 Mb
Total bases	943 Mb	946 Mb	946 Mb	946 Mb	1427 Mb

more abundant than retroelements (Chang et al. 2022); however, because so few elements are categorized for *L. rohita* (~24%), it is impossible to determine if this represents a lineage-specific difference or technical noise. Using these repeats, RepeatMasker masked 41.25% of the genome.

The annotation pipeline identified 51,079 primary transcripts, of which 31,274 survived the AED, gap, and overlapping filter criteria. BUSCO analysis shows the genome includes complete copies of 98.1% of the 3,640 orthologs in the actinopterygii_odb10

Table 3. BUSCO analysis for the genome and transcriptome, before and after AED filtering.

Type	Genome	Unfiltered transcriptome	Filtered transcriptome
Complete BUSCOs (C)	3571	3139	3078
Complete and single-copy BUSCOs (S)	3534	3064	3001
Complete and duplicated BUSCOs (D)	37	75	74
Fragmented BUSCOs (F)	23	192	170
Missing BUSCOs (M)	46	309	392
Total BUSCO groups searched	3640	3640	3640

database with 37 (1%) duplicated. The filtered transcriptome contains 84.5% of the total orthologs complete with 74 (2%) duplicated. An overview of the BUSCO analyses can be found in Table 3.

Comparative genomics

Our assembly (IGBB) was compared with the published and publicly available *L. rohita* assembly (CIFA), and annotated Cypriniformes assemblies from NCBI that were scaffold level or higher. Both the scaffold N50 and maximum length of the IGBB assembly are 30 Mb longer than the CIFA assembly (Table 2). The length distributions (Supplementary Fig. 2) show a similar separation, with overall greater contiguity in the IGBB genome. Interestingly, when the two *L. rohita* assemblies were pairwise aligned and plotted (Fig. 2), the CIFA assembly shows a few large gaps, specifically in Chr09 and Chr19, despite being larger in size. Due to the twofold size difference between the assemblies and the fragmentation of the CIFA assembly, the inverse comparison (i.e. IGBB aligned to CIFA) was not informative. Dot-plot alignments of the chromosome level Cypriniformes assemblies (Supplementary Fig. 3) generally exhibited similar chromosome structures, with some duplications and/or rearrangements. The assemblies for *Danio rerio* and *Triplophysa tibetana* were removed from the dot-plot grid since very few of the alignments passed the graphing threshold. Comparing the BUSCO results for the *L. rohita* assemblies, the IGBB assembly had fewer duplicate, fragmented, and missing BUSCOs than the CIFA assembly. Furthermore,

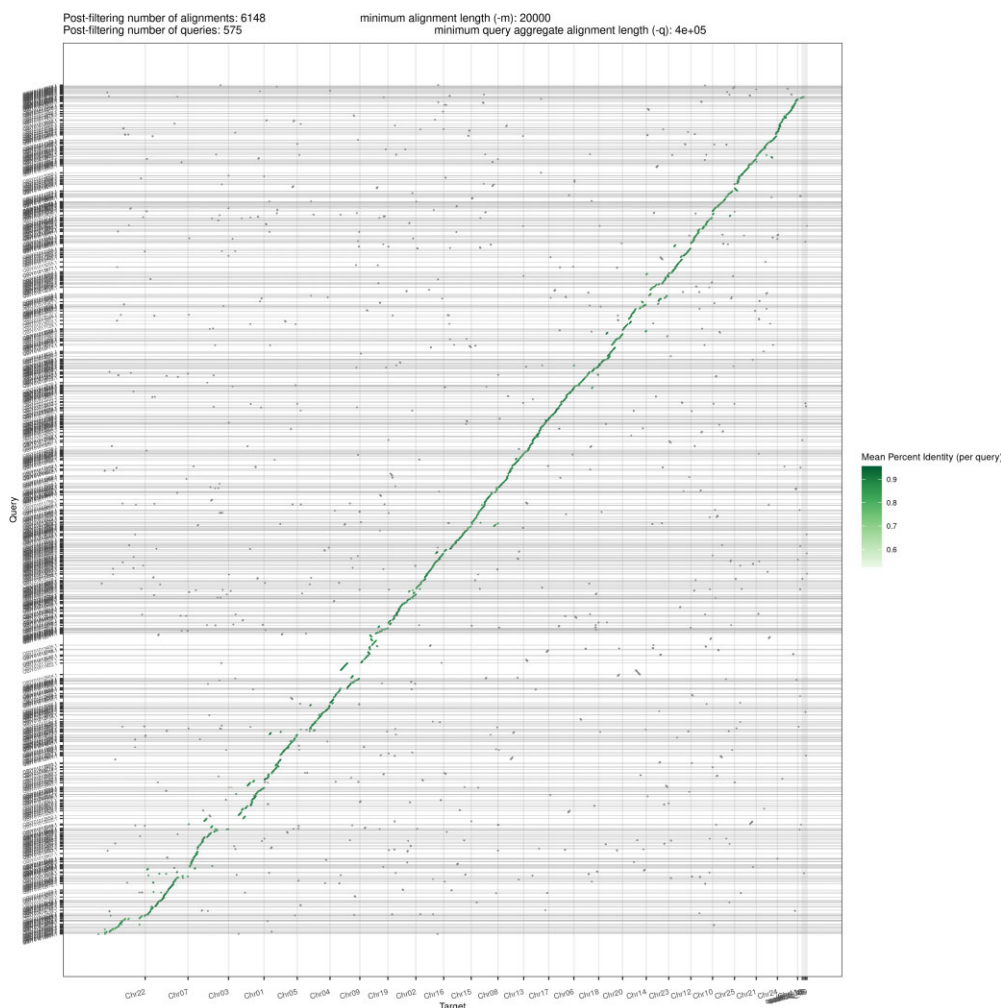


Fig. 2. Dot-plot between CIFA (y-axis) and IGBB (x-axis) *L. rohita* genomes, plotted using pafCoordsDotPlotly (<https://github.com/tpoorten/dotPlotly>).

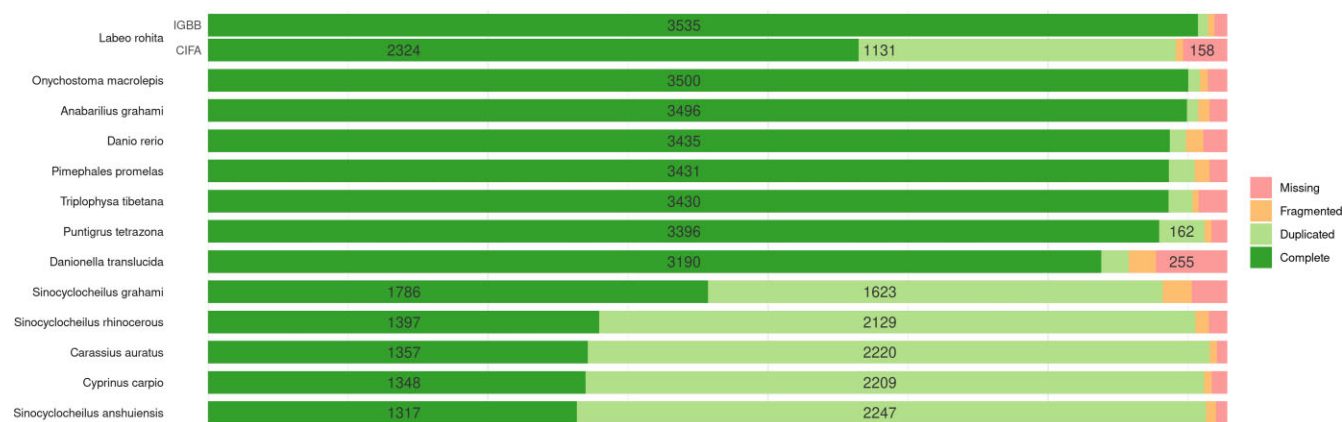


Fig. 3. BUSCO results for both *L. rohita* genomes (IGBB and CIFA) and the other included Cypriniformes genomes. The results for the two groups are sorted by complete single-copy BUSCOs.

Table 4. Summary statistics per species from OrthoFinder.

	Number of genes	Genes in orthogroups	Unassigned Genes	Orthogroups containing species	Species-specific orthogroups	Genes in species-specific orthogroups
<i>Anabarrilius grahami</i>	23,906	21,942 (91.8%)	1964 (8.2%)	15,217 (51.5%)	180	888 (3.7%)
<i>Carassius auratus</i>	96,703	93,988 (97.2%)	2715 (2.8%)	21,896 (74.2%)	472	1775 (1.8%)
<i>Cyprinus carpio</i>	80,686	78,684 (97.5%)	2002 (2.5%)	21,584 (73.1%)	287	964 (1.2%)
<i>Danio rerio</i>	52,829	51,951 (98.3%)	878 (1.7%)	20,671 (70.0%)	392	2184 (4.1%)
<i>Danionella translucida</i>	35,381	32,943 (93.1%)	2,438 (6.9%)	19,073 (64.6%)	540	1941 (5.5%)
<i>Labeo rohita</i>	31,274	29,904 (95.6%)	1370 (4.4%)	18,740 (63.5%)	161	1581 (5.1%)
<i>Onychostoma macrolepis</i>	24,754	24,483 (98.9%)	271 (1.1%)	19,276 (65.3%)	137	603 (2.4%)
<i>Pimephales promelas</i>	47,578	45,412 (95.4%)	2166 (4.6%)	19,884 (67.4%)	506	1826 (3.8%)
<i>Puntigrus tetrazona</i>	48,681	48,094 (98.8%)	587 (1.2%)	20,582 (69.7%)	129	517 (1.1%)
<i>Sinocyclocheilus anshuiensis</i>	68,474	66,456 (97.1%)	2018 (2.9%)	21,485 (72.8%)	114	344 (0.5%)
<i>Sinocyclocheilus grahami</i>	67,410	63,316 (93.9%)	4094 (6.1%)	22,326 (75.6%)	338	793 (1.2%)
<i>Sinocyclocheilus rhinoceros</i>	68,562	65,831 (96.0%)	2731 (4.0%)	21,884 (74.1%)	172	414 (0.6%)
<i>Triplophysa tibetana</i>	24,310	23,279 (95.8%)	1031 (4.2%)	18,734 (63.5%)	125	480 (2.0%)

the IGBB assembly had the most single-copy BUSCOs of any Cypriniformes (Fig. 3), even surpassing the model fish *D. rerio*. Notably, *Carassius auratus* and *Cyprinus carpio* are both allotetraploid fishes (Xu et al. 2019; Braasch 2020) and therefore exhibit a good deal of duplication in the dot-plots and BUSCO results. Lastly, the annotations for the Cypriniformes were compared using OrthoFinder. Of the 31,274 genes annotated, 29,904 (95.6%) were placed into 18,740 orthogroups, which comprise 63.5% of the total orthogroups found. Table 4 contains the summary statistics for all species used in the OrthoFinder analysis.

SNP discovery and population similarities among *L. rohita* fisheries

Aquaculture is an agricultural growth industry, producing 46% of the fish consumed worldwide. Over 50 million tonnes of finfish are raised in aquaculture each year, with the vast majority of aquaculture occurring in Asia (FAO 2020). Farm-raised *L. rohita* comprises 3.7% of the finfishes produced annually and represents the 11th most commonly farmed finfish (FAO 2020). Consumer preferences have been surveyed, identifying traits (e.g. length and weight) to prioritize in improvement programs (Mehar et al.

2022) along with disease resistance, some of which may be multigenic and complex. Genetically improved *L. rohita* seed is increasingly available to farmers (Das Mahapatra et al. 2007; Rasal et al. 2017; Hamilton 2019; Hamilton et al. 2019; SZA 2021; Hamilton et al. 2022); however, there is interest in further improving the characteristics of farmed *L. rohita*. Here we used ddRAD-sequencing in conjunction with the reference genome to provide insight into diversity and divergence among *L. rohita* in the Halda, Jamuna, and Padma river systems.

Patterns of divergence between the river systems (Fig. 4a and Supplementary Table 6, Supplementary Fig. 4a, Supplementary Fig. 5a) suggest that the geographically proximal Padma and Jamuna river systems (the Jamuna flows into the Padma) exhibited far less differentiation than either does to the hydrologically isolated and geographically distant Halda river system. While this pattern is similar to what was observed with silicoDArT markers (Hamilton et al. 2019), the greater number of nuclear sites surveyed here (i.e. 1.4 million) suggests that the differentiation between fish inhabiting these river systems is somewhat greater than previously reported using <2,000 SNP sites (Supplementary Table 6). These results (i.e. low differentiation between Padma

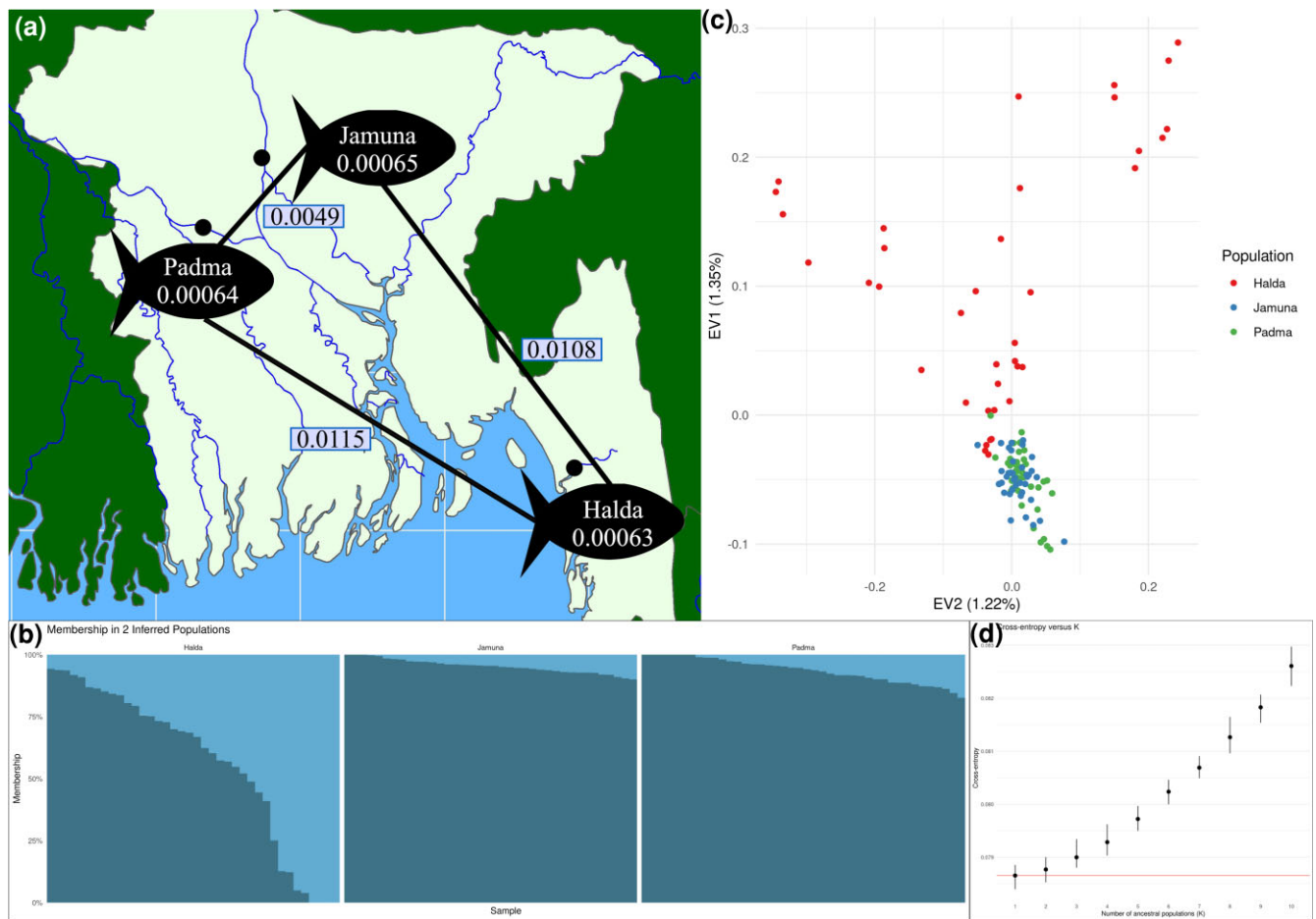


Fig. 4. a) Map of the river locations (dot), diversity (π) within each population (number within fish), and divergence (π_{xy}) between each population (number between fish); b) LEA predicted population structure ($k=2$) separated by river of origin. The vertical columns show the proportion (Q) assigned to each population for each individual; c) PCA plot of the filtered SNP set, colored by river of origin (D) Cross-entropy summary for the LEA analysis, using $K=1$ to 10 with 10 repetitions each.

and Jamuna and greater differentiation than previously reported) are congruent with an analysis of population structure ($k=2$) that reveals similar profiles for Padma- and Jamuna-based fish and a more divergent profile for fish from the Halda river system (Fig. 4b), which is reiterated in a principal component analysis (PCA) of fish from these rivers where fish from the Halda river system were adjacent to, but not intermingled with, fish from the Padma and Jamuna river systems (Fig. 4c). Interestingly, however, population structure analysis reaches optimization for these fish at $k=1$ (Fig. 4d), and the proportion of variation explained by the first two principal components is low ($\sim 1.5\%$ total), possibly indicating greater than expected admixture between Halda fish and those from the other two, geographically distant rivers (Fig. 4a).

Diversity among fish within each river was remarkably similar, ranging from 0.00063 in Halda to 0.00065 in Jamuna (π ; Supplementary Table 7, Supplementary Fig. 4b, Supplementary Fig. 5b). Notably, these estimates were nearly identical to the estimates of between-population divergence (π_{xy} ; Supplementary Table 8, Supplementary Fig. 4c, Supplementary Fig. 5c), which was 0.00064 for Padma-Halda and 0.00065 for both Padma-Jamuna and Jamuna-Halda, possibly indicating that these river populations are still representative of their shared ancestry. Diversity within populations (π) and divergence between (π_{xy}) populations were distributed relatively evenly across the chromosomes; however, in both cases, chromosomes 3, 4, and 22 were the only chromosomes that exhibited greater than average π

and π_{xy} , possibly indicating differences in selection and/or permeability on those chromosomes. Interestingly, while F_{st} for chromosomes 3 and 4 were not considerably different from many of the other chromosomes, chromosome 22 exhibited the greatest relative population divergence (F_{st} ; Supplementary Table 6), perhaps indicative of biologically relevant phenomena.

Sex-associated fragments

The genetics underlying SD in fish can be complicated and variable even within species (Devlin and Nagahama 2002; Volff et al. 2007; Parnell and Streebman 2013; Heule et al. 2014; Nguyen et al. 2021); however, controlling the sex ratio is essential to optimizing farming of finfish (Martinez et al. 2014). *L. rohita* breeding, for example, requires specific environmental conditions [i.e. monsoon (Qasim and Qayyum 1962; Natarajan and Jhingran 1963)], which is currently circumvented using hormonal induction (Bhattacharya 1999).

Despite its importance to aquaculture, the mechanisms governing SD in *L. rohita* are currently unknown. Karyotypic analyses suggest that if *L. rohita* has sex chromosomes, they are likely homomorphic (Bhatnagar et al. 2014), similar to many other fish (Heule et al. 2014), and are indistinguishable from the remaining chromosomes. We screened the *L. rohita* genome for regions linked to sex by evaluating read mapping in each ddRAD region from female vs male fish. Between 9.8 and 23.4 million (M) reads were uniquely mapped per sample to the 473,345 genomic regions

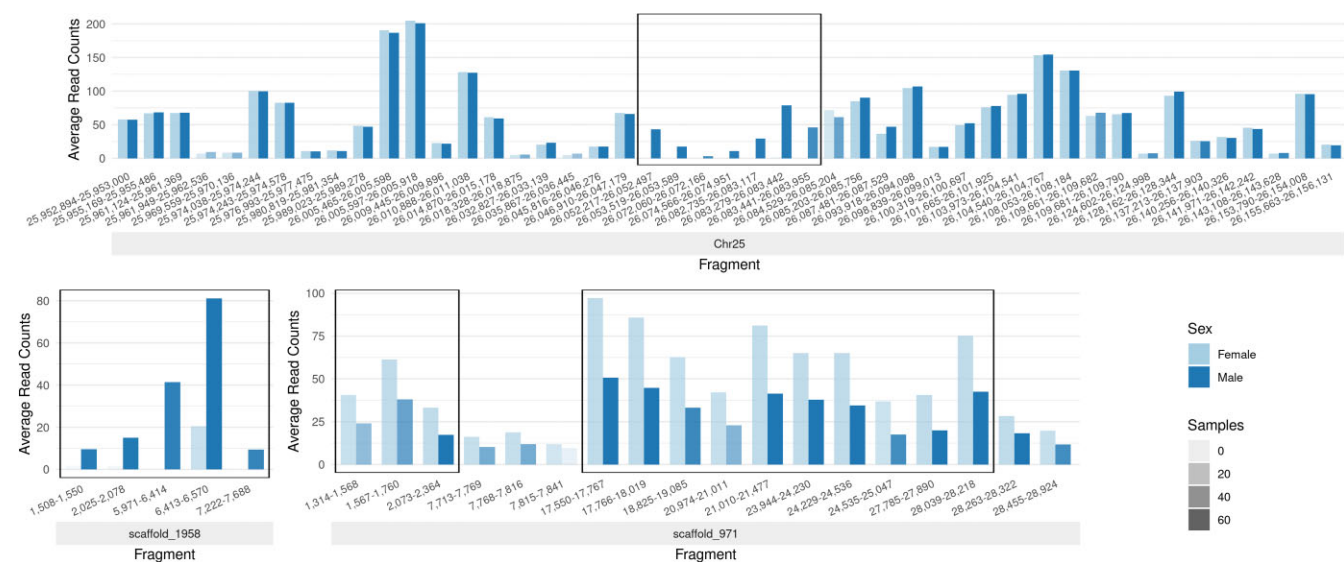


Fig. 5. Regions of statistically significant differences between male and female *L. rohita* read counts. Each pair shows the average read counts for male and female samples for a ddRAD fragment. Opacity of the bars vary with the number of samples present. The fragments are ordered but not spaced according to position. Contiguous statistically significant fragments are outlined.

occurring between the two restriction sites, as predicted by egads. Approximately 42% of these regions (200,543) had at least 50 samples with >50% of the region covered and were retained for two-sample Monte Carlo testing. Monte Carlo testing highlighted 25 fragments from three chromosomes/scaffolds (Chr25, scaffold_1958, and scaffold_971) as significantly (BH adj. P -value ≤ 0.05) different between females and males with respect to read coverage (Fig. 5 and Supplementary Table 9). Interestingly, the seven significant fragments on Chr25 are (1) contiguous, (2) cover approximately 30 kb (26,052,217–26,083,955), and (3) have no female samples mapping, suggesting that this may be a male-specific region of chromosome 25. The five fragments on scaffold_1958 show a similar pattern, albeit with a shorter total length (6.1 kb) and with female reads present, although significantly diminished, for a single region of the scaffold (~100 bp). Conversely, both male and female samples map to the 13 significant regions of scaffold_971 with reasonable coverage; however, the female samples generally had around double the mapping rate relative to the male samples, suggesting that this region may be represented by a greater copy number in females vs males. Together, these results suggest that *L. rohita* has a male-heterogametic (XX/XY) system of SD. Furthermore, since the sex chromosomes are indistinguishable by karyotype (Bhatnagar et al. 2014) and the uniquely male regions comprise only a small region of the chromosome, *L. rohita* may only have a Y-specific region (or “young Y”), similar to *Oryzias latipes* (Kondo et al. 2006); however, sequences similar to the *O. latipes* homolog for male-determination (i.e. dmY (Matsuda et al. 2002; Hornung et al. 2007)) were not found in the *L. rohita* assembly, indicating that further study is needed.

Conclusion

Despite its importance to aquaculture, *Labeo rohita* has only recently been studied using modern molecular techniques. Our flow cytometry, k-mer analysis, and NGS assembly of the *L. rohita* genome indicate a genome size of 0.97 Gb, a size 50–65% smaller than previously reported. Our IGBB reference-quality genome

for *L. rohita* both improves contiguity and removes the excessive redundancy of the previously existing CIFA draft genome sequence. The IGBB reference genome is a valuable resource for breeding programs and evolutionary biologists as demonstrated in our initial ddRAD-seq experiments. We find that, while fish from the connected rivers (Jamuna and Padma) are more similar in relative divergence, there remains a question of whether these populations are recently diverged from the Halda river system or if there remains some gene flow between all three, despite the hydrological and geographical isolation of the Halda. We also report candidate regions for SD in *L. rohita* that may underlie a male-heterogametic (XX/XY) system. While greater sampling will be required to understand the genetics underlying *L. rohita* SD and the population dynamics of these river systems, the present information provides a foundation for breeders to facilitate aquacultural improvement of *L. rohita*.

Data availability

The data used for the *Labeo rohita* genome and annotation are available at NCBI under the BioProject PRJNA650519. The assembled genome sequence and annotations are available at GenBank under accessions JACTAM000000000. The raw data is available at the SRA (Sequence Read Archive) under accessions SRR12580210–SRR12580221. The ddRAD-seq data used for SNP discovery, population analyses, and sex-associated fragment analysis are available under the BioProject PRJNA841581 and the SRA accessions SRR19358298 – SRR19358417. The RepeatModeler and RepeatMasker analysis output, along with the unfiltered ddRAD vcf, are available at <https://doi.org/10.5281/zenodo.7377776>.

Acknowledgements

The authors thank the Iowa State University Flow Cytometry Facility and ResearchIT unit for technical and computational support, respectively, WorldFish Carp Genetic Improvement Program

staff based in Jashore, Bangladesh for managing and sampling fish, and Mahirah Mahmuddin for sample collation and management.

Funding

The authors declare that this work was in part supported through the United States Agency for International Development (USAID) “Innovate4Fish Feed the Future Fish Innovation Lab- Quick Start” (Grant ID 7200AA18CA00030). Fin clip samples were collected with funding from the United States Agency for International Development (USAID) Aquaculture for Income and Nutrition project (Grant ID EEM-G-00-04-00013-00).

Conflicts of interest

The authors declare no conflict of interest.

[Supplemental material](#) available at G3 online.

Literature cited

- Alonge M, Lebeigle L, Kirsche M, Aganezov S, Wang X, Lippman ZB, Schatz MC, Soyk S. Automated assembly scaffolding elevates a new tomato system for high-throughput genome editing. *Genome Biol* 2022;23(1):258. doi:10.1101/2021.11.18.469135.
- Bhatnagar A, Yadav AS, Kamboj K. Karyomorphology of three Indian Major carps from Haryana, India. *J Fish*. 2014;8(2):95–103. doi:10.3153/jfscom.201413.
- Bhattacharya S. Recent advances in the hormonal regulation of gonadal maturation and spawning in fish. *Curr Sci*. 1999;76(3):342–349.
- Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114–2120. doi:10.1093/bioinformatics/btu170.
- Borodovsky M, Lomsadze A. Eukaryotic gene prediction using GeneMark.hmm-E and GeneMark-ES. *Curr Protoc Bioinforma*. 2011;35(1):1. doi:10.1002/0471250953.bi0406s35.
- Braasch I. Genome evolution: domestication of the allopolyploid goldfish. *Curr Biol*. 2020;30(14):R812–R815. doi:10.1016/j.cub.2020.05.073.
- Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics Bioinforma*. 2021;3(1):lqaa108. doi:10.1093/nargab/lqaa108.
- Budd AM, Banh QQ, Domingos JA, Jerry DR. Sex control in fish: approaches. Challenges and opportunities for aquaculture. *J Mar Sci Eng*. 2015;3(2):329–355. doi:10.3390/jmse3020329.
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. *BMC Bioinformatics*. 2009;10(1):421. doi:10.1186/1471-2105-10-421.
- Chang N-C, Rovira Q, Wells J, Feschotte C, Vaquerizas JM. Zebrafish transposable elements show extensive diversification in age, genomic distribution, and developmental expression. *Genome Res*. 2022;32(7):1408–1423. doi:10.1101/gr.275655.121.
- Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. Twelve years of SAMtools and BCFtools. *GigaScience*. 2021;10(2):giab008. doi:10.1093/gigascience/giab008.
- Das P, Sahoo L, Das SP, Bit A, Joshi CG, Kushwaha B, Kumar D, Shah TM, Hinsu AT, Patel N, et al. De novo assembly and genome-wide SNP discovery in Rohu Carp, *Labeo rohita*. *Front Genet*. 2020;11:386. doi:10.3389/fgene.2020.00386.
- Das Mahapatra K, Saha J, Sarangi N, Jana R, Gjerde B, Nguyen N, Khaw HL, Ponzoni R. Genetic improvement and dissemination of rohu (*Labeo rohita*, Ham.) in India. Proceedings of Seventeenth Conference Association for the Advancement of Animal Breeding and Genetics, 23rd–26th September 17; 2007
- De Coster W, D’Hert S, Schultz DT, Cruts M, Van Broeckhoven C. Nanopack: visualizing and processing long-read sequencing data. *Bioinformatics*. 2018;34(15):2666–2669. doi:10.1093/bioinformatics/bty149.
- Devlin RH, Nagahama Y. Sex determination and sex differentiation in fish: an overview of genetic, physiological, and environmental influences. *Aquaculture*. 2002;208(3–4):191–364. doi:10.1016/S0044-8486(02)00057-1.
- DoF. Yearbook of Fisheries Statistics of Bangladesh, 2019–20. Fisheries Resources Survey System (FRSS), Department of Fisheries. Bangladesh: Ministry of Fisheries and Livestock; 2020.
- Doyle JJ, Doyle JL, editors. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem Bull*. 1987
- Emms DM, Kelly S. Orthofinder: phylogenetic orthology inference for comparative genomics. *Genome Biol*. 2019;20(1):238. doi:10.1186/s13059-019-1832-y.
- FAO. The State of World Fisheries and Aquaculture 2020: Sustainability in Action. Rome, Italy: FAO; 2020 (The State of World Fisheries and Aquaculture (SOFIA)). [accessed 2022 May 11]. <https://www.fao.org/documents/card/en/c/ca9229en/>.
- Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. Repeatmodeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A*. 2020;117(17):9451–9457. doi:10.1073/pnas.1921046117.
- Frichot E, François O. LEA: an R package for landscape and ecological association studies. *Methods in Ecology and Evolution*. 2015;6(8):925–929. doi:10.1111/2041-210X.12382.
- Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, Phillippy AM, Koren S. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS Comput Biol*. 2019;15(8):e1007273. doi:10.1371/journal.pcbi.1007273.
- Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*. 2011;29(7):644–652. doi:10.1038/nbt.1883.
- Greilhuber J. Intraspecific variation in genome size in angiosperms: identifying its existence. *Ann Bot*. 2005;95(1):91–98. doi:10.1093/aob/mci004.
- Hamilton MG. Management of Inbreeding in carp hatcheries in Myanmar; 2019. [accessed 2022 May 17]. <https://digitalarchive.worldfishcenter.org/handle/20.500.12348/3859>.
- Hamilton MG, Mekki W, Kilian A, Benzie JAH. Single nucleotide polymorphisms (SNPs) reveal sibship among founders of a Bangladeshi Rohu (*Labeo rohita*) breeding population. *Front Genet*. 2019;10:597. doi:10.3389/fgene.2019.00597.
- Hamilton MG, Mekki W, MdB A, Benzie JAH. Early selection to enhance genetic gain in a rohu (*Labeo rohita*) genetic improvement program. *Aquaculture*. 2022;553:738058. doi:10.1016/j.aquaculture.2022.738058.
- Heule C, Salzburger W, Böhne A. Genetics of sexual development: an evolutionary playground for fish. *Genetics*. 2014;196(3):579–591. doi:10.1534/genetics.114.161158.
- Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*. 2011;12(1):491. doi:10.1186/1471-2105-12-491.

- Hornung U, Herpin A, Schartl M. Expression of the male determining gene *dmrt1bY* and its autosomal coorthologue *dmrt1a* in medaka. *Sex Dev.* 2007;1(3):197–206. doi:10.1159/000102108.
- Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics.* 2010;11(1):119. doi:10.1186/1471-2105-11-119.
- Jackman SD, Vandervalk BP, Mohamadi H, Chu J, Yeo S, Hammond SA, Jahesh G, Khan H, Coombe L, Warren RL, et al. ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. *Genome Res.* 2017;27(5):768–777. doi:10.1101/gr.214346.116.
- Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al. Interproscan 5: genome-scale protein function classification. *Bioinformatics.* 2014;30(9):1236–1240. doi:10.1093/bioinformatics/btu031.
- Kendig KI, Baheti S, Bockol MA, Drucker TM, Hart SN, Heldenbrand JR, Hernaez M, Hudson ME, Kalmbach MT, Klee EW, et al. Sentieon DNaseq variant calling workflow demonstrates strong computational performance and accuracy. *Front Genet.* 2019; 10. doi:10.3389/fgene.2019.00736
- Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol.* 2019;37(8):907–915. doi:10.1038/s41587-019-0201-4.
- Kondo M, Hornung U, Nanda I, Imai S, Sasaki T, Shimizu A, Asakawa S, Hori H, Schmid M, Shimizu N, et al. Genomic organization of the sex-determining and adjacent regions of the sex chromosomes of medaka. *Genome Res.* 2006;16(7):815–826. doi:10.1101/gr.5016106.
- Korunes KL, Samuk K. paxy: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Mol Ecol Resour.* 2021;21(4):1359–1368. doi:10.1111/1755-0998.13326.
- Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* 2019;20(1):278. doi:10.1186/s13059-019-1910-1.
- Li H. Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM; 2013 May 26. ArXiv13033997 Q-Bio. [accessed 2022 Mar 3]. <http://arxiv.org/abs/1303.3997>.
- Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* 2018;34(18):3094–3100. doi:10.1093/bioinformatics/bty191.
- Magbanua ZV, Hsu C, Pechanova O, Arick M, Grover CE, Peterson DG. 2022. Innovations in double digest restriction-site associated DNA sequencing (ddRAD-Seq) method for more efficient SNP identification. *Anal Biochem.* 2023;662:115001. doi: 10.1016/j.ab.2022.115001
- Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO Update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol.* 2021;38(10):4647–4654. doi:10.1093/molbev/msab199.
- Mapleson D, Venturini L, Kaithakottil G, Swarbreck D. Efficient and accurate detection of splice junctions from RNA-seq with Portcullis. *GigaScience.* 2018;7(12):12. doi:10.1093/gigascience/giy131.
- Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics.* 2011;27(6): 764–770. doi:10.1093/bioinformatics/btr011.
- Martínez P, Viñas AM, Sánchez L, Díaz N, Ribas L, Piferrer F. Genetic architecture of sex determination in fish: applications to sex ratio control in aquaculture. *Front Genet.* 2014;5:340. doi: 10.3389/fgene.2014.00340
- Matsuda M, Nagahama Y, Shinomiya A, Sato T, Matsuda C, Kobayashi T, Morrey CE, Shibata N, Asakawa S, Shimizu N, et al. DMY Is a Y-specific DM-domain gene required for male development in the medaka fish. *Nature.* 2002;417(6888):559–563. doi: 10.1038/nature751.
- Mehar M, Mekki W, Mcdougall C, Benzie JAH. Preferences for rohu fish (*L. rohita*) traits of women and men from farming households in Bangladesh and India. *Aquaculture.* 2022;547: 737480. doi:10.1016/j.aquaculture.2021.737480.
- Natarajan A, Jhingran A. On the biology of Catla catla (Ham.) from the river Jamuna. *Proc Nat Inst Sci India.* 1963;29:326–355.
- Nguyen DHM, Panthum T, Ponjarat J, Laopichienpong N, Kraichak E, Singchat W, Ahmad SF, Muangmai N, Peyachoknagul S, Na-Nakorn U, et al. An investigation of ZZ/ZW and XX/XY sex determination systems in North African Catfish (*Clarias gariepinus*, Burchell, 1822). *Front Genet.* 2021;11:562856. doi:10.3389/fgene.2020.562856.
- Parnell NF, Streelman JT. Genetic interactions controlling sex and color establish the potential for sexual conflict in Lake Malawi cichlid fishes. *Heredity (Edinb).* 2013;110(3):239–246. doi:10.1038/hdy.2012.73.
- Patel A, Das P, Barat A, Sarangi N. Estimation of genome size in Indian major carps *Labeo rohita* (Hamilton), *Catla catla* (Hamilton), *Cirrhinus mrigala* (Hamilton) and *Labeo calbasu* (Hamilton) by Feulgen microdensitometry method. *Ind J Fish.* 2009;56(1):65–67.
- Pellicer J, Leitch IJ. The application of flow cytometry for estimating genome size and ploidy level in plants. *Methods Mol Biol.* 2014; 1115:279–307. doi:10.1007/978-1-62703-767-9_14.
- Qasim SZ, Qayyum A. Spawning frequencies and breeding seasons of some freshwater fishes with special reference to those occurring in the plains of northern India. *Ind J Fish.* 1962;8(1):24–43.
- Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* 2010;26(6):841–842. doi:10.1093/bioinformatics/btq033.
- Rasal KD, Chakrapani V, Pandey AK, Rasal AR, Sundaray JK, Ninawe A, Jayasankar P. Status and future perspectives of single nucleotide polymorphisms (SNPs) markers in farmed fishes: way ahead using next generation sequencing. *Gene Rep.* 2017;6:81–86. doi: 10.1016/j.genrep.2016.12.004.
- Rasal KD, Iquebal MA, Pandey A, Behera P, Jaiswal S, Vasam M, Dixit S, Raza M, Sahoo L, Nandi S, et al. Revealing liver specific microRNAs linked with carbohydrate metabolism of farmed carp, *Labeo rohita* (Hamilton, 1822). *Genomics.* 2020;112(1):32–44. doi:10.1016/j.ygeno.2019.07.010.
- Rasal KD, Sundaray JK. Status of genetic and genomic approaches for delineating biological information and improving aquaculture production of farmed rohu, *Labeo rohita* (Ham, 1822). *Rev Aquac.* 2020;12(4):2466–2480. doi:10.1111/raq.12444.
- R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2020. <https://www.R-project.org/>.
- Robinson N, Baranski M, Mahapatra K, Saha J, Das S, Mishra J, Das P, Kent M, Arnyasi M, Sahoo P. A linkage map of transcribed single nucleotide polymorphisms in rohu (*Labeo rohita*) and QTL associated with resistance to *Aeromonas hydrophila*. *BMC Genomics.* 2014;15(1):541. doi:10.1186/1471-2164-15-541.
- Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. *Nat Methods.* 2020;17(2):155–158. doi:10.1038/s41592-019-0669-3.
- Sahoo L, Das P, Sahoo B, Das G, Meher PK, Udit UK, Mahapatra KD, Sundaray JK. The draft genome of *Labeo catla*. *BMC Res Notes.* 2020;13(1):411. doi:10.1186/s13104-020-05240-w.
- Sahoo L, Meher PK, Das Mahapatra K, Saha JN, Jayasankar P, Das P. A molecular tool for parentage analysis in Indian major carp, *Labeo rohita* (Hamilton, 1822). *Aquac Int.* 2017;25(3):1159–1166. doi:10.1007/s10499-016-0104-z.

- Sahoo L, Sahoo S, Mohanty M, Sankar M, Dixit S, Das P, Rasal KD, Rather MA, Sundaray JK. Molecular characterization, computational analysis and expression profiling of *Dmrt1* gene in Indian major carp, *Labeo rohita* (Hamilton 1822). *Anim Biotechnol.* 2021; 32(4):413–426. doi:10.1080/10495398.2019.1707683.
- Sahu DK, Panda SP, Panda S, Das P, Meher PK, Hazra RK, Peatman E, Liu ZJ, Eknath AE, Nandi S. Identification of reproduction-related genes and SSR-markers through expressed sequence tags analysis of a monsoon breeding carp rohu, *Labeo rohita* (Hamilton). *Gene.* 2013;524(1):1–14. doi:10.1016/j.gene.2013.03.111.
- Smit A, Hubley R, Green P. RepeatMasker Open-4.0; 2013. <http://www.repeatmasker.org>.
- Stanke M, Schöffmann O, Morgenstern B, Waack S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics.* 2006;7(1):62. doi:10.1186/1471-2105-7-62.
- Sun L, Gao T, Wang F, Qin Z, Yan L, Tao W, Li M, Jin C, Ma L, Kocher TD, et al. Chromosome-level genome assembly of a cyprinid fish *Onychostoma macrolepis* by integration of nanopore sequencing, Bionano and Hi-C technology. *Mol Ecol Resour.* 2020;20(5):1361–1371. doi:10.1111/1755-0998.13190.
- SZA. “Subarna Ruhi” developed after a decade of efforts; 2021. *Obs Online Rep.* [accessed 2022 May 19]. <https://www.observerbd.com/details.php?id=316542>.
- Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and cufflinks. *Nat Protoc.* 2012;7(3):562–578. doi:10.1038/nprot.2012.016.
- The UniProt Consortium. Uniprot: the universal protein knowledge-base in 2021. *Nucleic Acids Res.* 2021;49(D1):D480–D489. doi:10.1093/nar/gkaa1100.
- Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* 2017;27(5):737–746. doi:10.1101/gr.214270.116.
- Venturini L, Caim S, Kaithakottil GG, Mapleson DL, Swarbreck D. Leveraging multiple transcriptome assembly methods for improved gene structure annotation. *GigaScience.* 2018;7(8):8. doi:10.1093/gigascience/giy093.
- Volff J-N, Nanda I, Schmid M, Schartl M. Governing sex determination in fish: regulatory putsches and ephemeral dictators. *Sex Dev.* 2007;1(2):85–99. doi:10.1159/000100030.
- Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC. Genomescope: fast reference-free genome profiling from short reads. *Bioinformatics.* 2017;33(14):2202–2204. doi:10.1093/bioinformatics/btx153.
- Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One.* 2014;9(11):e112963. doi:10.1371/journal.pone.0112963.
- Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, Grolemund G, Hayes A, Henry L, Hester J, et al. Welcome to the Tidyverse. *J Open Source Softw.* 2019;4(43):1686. doi:10.21105/joss.01686.
- Xu P, Xu J, Liu G, Chen L, Zhou Z, Peng W, Jiang Y, Zhao Z, Jia Z, Sun Y, et al. The allotetraploid origin and asymmetrical genome evolution of the common carp *Cyprinus carpio*. *Nat Commun.* 2019; 10(1):4625. doi:10.1038/s41467-019-12644-1.
- Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics.* 2012;28(24):3326–3328. doi:10.1093/bioinformatics/bts606.

Communicating editor: A. Whitehead