



RESEARCH ARTICLE

REVISED Beyond the Beauty: Meristic and Genomic Signatures of**Bangka's Endemic Betta Fishes**

[version 2; peer review: 2 approved, 1 approved with reservations]

Helmizuryani Helmizuryani¹, Saleh Hidayat², Muhammad Nizar ¹,
 Andhika Prima Prasetyo ³, Bobby Muslimin ³, Mochammad Zamroni ⁴,
 Dessy Nurul Astuti⁴, Swarlanda Swarlanda⁵, Destra Ramadhanu⁵,
 Luthfi Nurhidayat ⁶

¹Study Program of Aquaculture, Faculty of Agriculture, Universitas Muhammadiyah Palembang, Palembang, South Sumatra, 30263, Indonesia

²Study Program of Biology Education, Faculty of Teacher Training and Education, Universitas Muhammadiyah Palembang, Palembang, South Sumatra, 30263, Indonesia

³Research Center for Applied Zoology, National Research and Innovation Agency, Cibinong, West Java, 16911, Indonesia

⁴Research Center for Biota System, National Research and Innovation Agency, Cibinong, West Java, 16911, Indonesia

⁵Yayasan Ikan Endemik Bangka Belitung, Pangkal Pinang, Bangka Belitung Island, 33123, Indonesia

⁶Faculty of Biology, Universitas Gadjah Mada, Sleman, Special Region of Yogyakarta, 55281, Indonesia

v2 First published: 10 Feb 2026, 15:226
<https://doi.org/10.12688/f1000research.174779.1>

Latest published: 01 Jun 2026, 15:226
<https://doi.org/10.12688/f1000research.174779.2>

Abstract**Background**

The genus *Betta* (family Osphronemidae) comprises over 70 species, many of which are endemic to Southeast Asia and highly vulnerable to habitat loss. While *Betta splendens* is well studied due to its importance in the ornamental fish trade, most wild *Betta* species remain poorly characterized, particularly at the genomic level. The Bangka Islands of Indonesia harbor several endemic *Betta* species threatened by peatland degradation.

Methods

We conducted an integrated meristic and genomic comparison of three endemic Bangka Island *Betta* species—*Betta burdigala*, *B. chloropharynx*, and *B. schalleri*. Specimens were collected from peatland waters in Bangka, and meristic traits were examined to confirm diagnostic characteristics. High-molecular-weight DNA was extracted and sequenced using Oxford Nanopore PromethION technology, followed by *de novo* assembly and reference-guided

Open Peer Review**Approval Status** ✓ ? ✓

	1	2	3
version 2			
(revision)	✓		✓
01 Jun 2026	view		view
	↑		
version 1	?	?	
10 Feb 2026	view	view	

1. **Bingpeng Xing**, Third Institute of Oceanography Ministry of Natural Resources, Xiamen, China
2. **Nicholas S Kron** , University of Miami, Rickenbacker Cswy, USA
Genevia Technologies, Finland, Finland
3. **Axel Janke**, Senckenberg Biodiversity and Climate Research Centre, Frankfurt am Main, Germany

scaffolding using the *Betta splendens* genome.

Results

The meristic analysis confirmed features consistent with their taxonomic placement within the *coccina*, *waseri*, and *pugnax* groups. Genome assemblies were highly contiguous and complete (BUSCO >97%), with *B. chloropharynx* showing the largest genome size, highest scaffold N50, and elevated retrotransposon content. Gene duplication analysis revealed dispersed duplications as the dominant category across all genomes, with variation in tandem and proximal duplicates. Comparative genomic analysis demonstrated conserved collinearity, with *B. chloropharynx* and *B. schalleri* showing the closest relationship, while *B. burdigala* diverged earlier. The Colony Stimulating Factor 1 Receptor A (CSF1RA) protein phylogenetic tree closely resembles the phylogenetic tree of nine fish species based on NCBI taxonomic data. We also identified two massive protein insertions in the CSF1RA of *B. burdigala*.

Conclusions

This study provides morphological and genomic evidence supporting the distinctiveness of Bangka's endemic *Betta* species and delivers essential genomic resources for evolutionary research and conservation of these endangered freshwater fishes.

Keywords

Betta burdigala, *Betta chloropharynx*, *Betta schalleri*, meristic analysis, genome assembly, synteny, conservation genomics, Bangka Island

Any reports and responses or comments on the article can be found at the end of the article.



This article is included in the **Genomics and Genetics** gateway.

Corresponding author: Luthfi Nurhidayat (luthfibio@ugm.ac.id)

Author roles: **Helmizuryani H:** Conceptualization, Data Curation, Investigation, Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Hidayat S:** Data Curation, Investigation, Methodology, Validation; **Nizar M:** Data Curation, Investigation, Methodology, Validation; **Prasetyo AP:** Investigation, Validation; **Muslimin B:** Data Curation, Formal Analysis, Investigation, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Zamroni M:** Data Curation, Investigation, Methodology, Validation; **Astuti DN:** Data Curation, Investigation, Validation; **Swarlanda S:** Data Curation, Investigation, Methodology, Validation; **Ramadhanu D:** Data Curation, Investigation, Methodology, Validation; **Nurhidayat L:** Data Curation, Formal Analysis, Software, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: 123/C3/DT.05.00/PL/2025

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2026 Helmizuryani H *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Helmizuryani H, Hidayat S, Nizar M *et al.* **Beyond the Beauty: Meristic and Genomic Signatures of Bangka's Endemic Betta Fishes [version 2; peer review: 2 approved, 1 approved with reservations]** F1000Research 2026, 15:226 <https://doi.org/10.12688/f1000research.174779.2>

First published: 10 Feb 2026, 15:226 <https://doi.org/10.12688/f1000research.174779.1>

REVISED Amendments from Version 1

In this revised manuscript, we have addressed all reviewer concerns by moderating our evolutionary interpretations and replacing population genetic assertions, such as purifying selection or relaxed constraint, with more conservative language emphasizing sequence conservation and structural variation, which more appropriately reflects the use of single-individual reference assemblies. We also explicitly acknowledged the potential structural biases associated with reference-guided scaffolding in both the Methods and Discussion sections, clarifying that our collinearity analyses primarily reflect shared gene order and syntenic patterns rather than definitive chromosome-scale structural identity. In addition, we improved methodological transparency by incorporating contamination screening workflows using Porechop_ABI and NCBI FCS-GX, as well as by adding sequencing coverage estimates and read N50 statistics to the Materials and Methods section. We also updated Figure 3 to include 21-mer-based genome size estimations.

To further support the novel CSF1RA insertions identified in *B. burdigala*, we mapped raw Oxford Nanopore long reads back to the assembled genome to verify stable coverage across the insertion locus and deposited the alignment evidence in the extended data repository on Figshare. This includes “Suppl. Fig. 3. Validation of the CSF1RA insertion using Oxford Nanopore long-read alignments” and “Supplementary File 4. Coverage depth of the CSF1RA_1 region based on Oxford Nanopore long-read alignments.” We also added “Supplementary File 5. QAST assembly quality assessment statistics for the three Betta genome assemblies” to the same repository.

Finally, we performed a comprehensive language and formatting revision throughout the manuscript, including standardization of taxonomic formatting, clarification of the single-gene nature and limitations of the CSF1RA phylogenetic analysis, and correction of background statements related to regional species diversity and IUCN conservation status terminology.

Any further responses from the reviewers can be found at the end of the article

Introduction

The genus *Betta* (family Osphronemidae) includes over 70 recognized species, many of which are endemic to Southeast Asia. Despite the global popularity of the domesticated *Betta splendens* in the ornamental fish trade, the majority of wild *Betta* species remain under-researched. Indonesia harbors nine wild *Betta* species classified as Critically Endangered (CN) and seven species categorized as Endangered (EN).^{1–3} In comparison, Thailand contains two Critically Endangered (CE) wild *Betta* species and one Endangered (EN) species.^{4,5} The high levels of endemism and restricted ranges of many *Betta* species make them particularly vulnerable to habitat loss and environmental changes.^{2,5}

The Bangka Islands, located off the east coast of Sumatra, are home to several unique *Betta* species adapted to specialized freshwater habitats such as peat swamps and slow-moving forest streams. Three *Betta* species—*B. burdigala*, *B. chloropharynx*, and *B. schalleri*—are endemic to the Bangka Islands.¹ According to IUCN Red List, the conservation status for them were Critically Endangered (CR) and Endangered (EN). The lack of molecular resources poses a barrier to implementing effective conservation strategies for these fishes. Previous studies on the endemic *Betta* species from Bangka Island have included genetic analyses based on mitochondrial DNA and eDNA metabarcoding, as well as comprehensive morphological examinations, all of which have contributed to clarifying the genetic relationships and taxonomic status of these species. However, the whole-genome evolutionary landscape of the endemic *Betta* species from Bangka Island has not yet been explored.⁶

Genome assemblies are indeed foundational resources for advancing biological research and conservation management. They provide critical insights into genetic diversity, evolutionary biology, and species conservation strategies. High-quality genomic data could be used for Long-read sequencing and high-throughput chromosome (Hi-C) technology. Long-read sequencing technologies are essential for producing high-quality genome assemblies. High-quality genomic data can be generated using long-read sequencing technologies such as Oxford Nanopore sequencing, which enables the production of highly contiguous genome assemblies. These technologies help resolve complex repeats and haplotype heterozygosity, which are sources of assembly errors.⁷ Genome assemblies could be informative evidence for scientific management decisions and tools for understanding the basis of genetic adaptation in various species.^{8,9}

Bangka Island presents a unique opportunity to investigate the genetic foundations of adaptation and diversity in insular freshwater fish species. The endemic *Betta* species of the island have likely undergone distinct evolutionary trajectories due to the island's unique environmental conditions and geographic isolation. Producing genome assemblies for these taxa will address a significant knowledge gap and serve as a foundational step for future conservation genomics efforts in the region. This study aims to generate preliminary genome assemblies for three endemic *Betta* species originating from Bangka, Indonesia, utilizing whole-genome sequencing methodologies.

Materials and methods

Ethics statement

This study received ethical approval from the Ethics Commission for Animal Care and Use, National Research and Innovation Agency Republic of Indonesia (Approval No. 220/KE.02/SK/09/2024). All animal procedures followed institutional regulations and adhered to the ARRIVE 2.0 reporting guidelines, with the corresponding checklists available at <https://doi.org/10.6084/m9.figshare.31136455>.¹⁰

Sample collection

This study examines three endemic species of Betta: *Betta burdigala*, *Betta chloropharynx*, and *Betta schalleri*. The fish samples were collected from peatland water in South Bangka, Bangka Island, Indonesia. The map showing the origin locations of specimens on Bangka Island is presented in Figure 1. The males of *B. schalleri* and *B. chloropharynx* have been found in Central Bangka (2°22'00.7"S 106°10'57.9"E) and *B. burdigala* has been found in South Bangka (2°50'24.8"S 106°26'12.7"E). The specimens were collected using a scope net with a mesh size of 0.5 mm as active gear. In-situ sampling data for morphometric and meristic analysis were obtained from three individuals each of *B. schalleri*, *B. chloropharynx*, and *B. burdigala*. The ten morphometric and seven meristic characters (Suppl Fig 1) based on Nur et al. (2022)² were measured using a digital caliper with an accuracy 0.1 mm. The sampling design in this study used purposive sampling based on mature females or males with different shapes and colors. Males have longer ventral fins than females.¹¹ The males also have more vibrant and diverse body colors than females.¹² The morphometric and meristic data were analyzed descriptively.

One male specimen from each Betta species was euthanized using the rapid chilling method, which involved transferring the fish to water maintained at approximately 2°C with crushed ice.¹³ The fish was left in this condition for 20 minutes until opercular movement ceased. Subsequently, tissue samples were collected, fixed with absolute ethanol, and stored in cryotubes for whole genome sequencing at the Central Sequencing Laboratory, BRIN, Bogor, Indonesia.

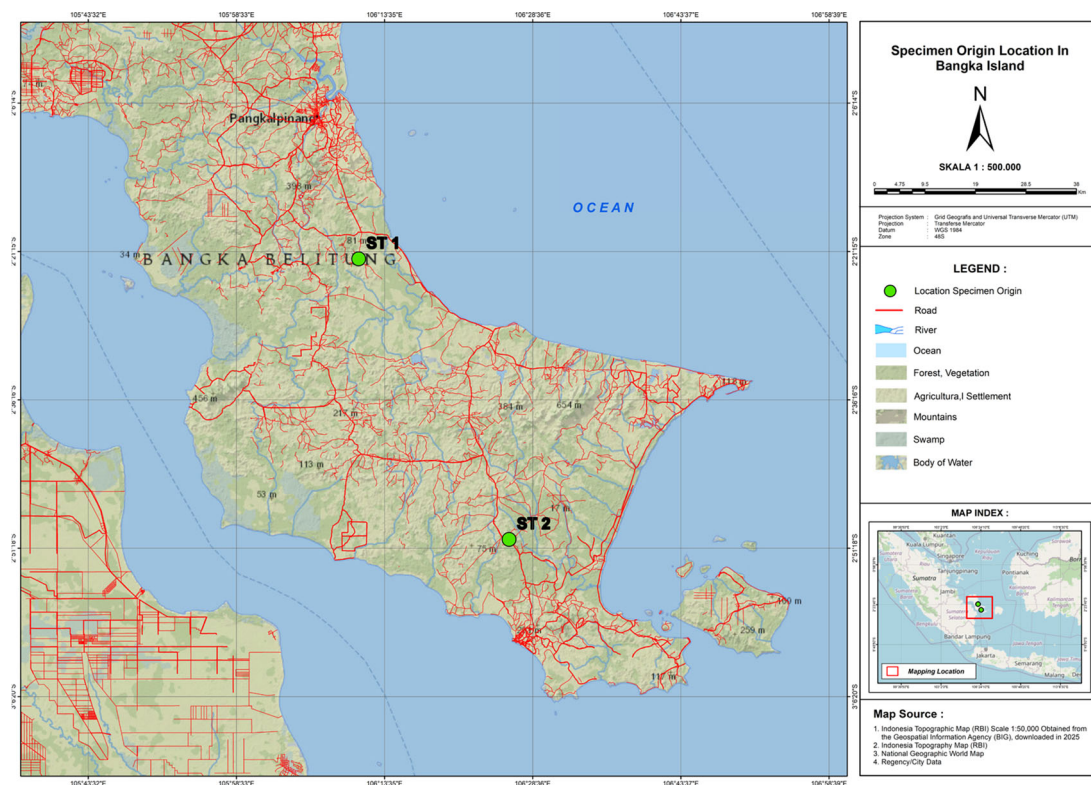


Figure 1. The specimen origin locations of *Betta schalleri* and *Betta chloropharynx* are found in ST.1 (Central Bangka), while *Betta burdigala* is found in ST.2 (South Bangka).

Genomic DNA extraction and sequencing

The extraction of genomic DNA was performed using the Applied Biosystem MagMax™ DNA Multi-Sample 2.0 kit (Thermo Fisher Scientific; CAT. A36570) following the manufacturer's instructions. Twenty mg of fish tissue is used for genome DNA extraction material for all *Betta* specimens. The genome concentration from the extraction was checked using Nanodrop and Qubit. Genome concentration by Nanodrop value between 227.45–289 ng/μl and Qubit value between 89.4–91.6 (ng/μl). The genome quality was checked using gel agarose with TBE agarose concentration 1%. The genomic DNA was subsequently prepared for library construction using the Ligation Sequencing DNA V14 kit (Oxford Nanopore Technology; SQK-LSK114) in accordance with the manufacturer's instructions. Following this, the library was sequenced on a PromethION device utilizing PromethION Flow Cells Packs (Oxford Nanopore Technology; FLO-PRO114M). The sequencing parameters we employed included a run duration of 96 hours, a pre-scan interval of 1.5 hours, basecalling using the High-accuracy model at 400 bps, and a minimum Q score of 9. The sequencing software utilizes MinKNOW (25.03.7), Bream (8.4.4), configuration (6.4.10), basecalling Dorado (dna_r10.4.1_e8.2_400bps_hac@4.3.0), and MinKNOW Core (6.4.8). The genome sequence data for *B. burdigala* comprised 16.38 million reads totaling 37.39 Gb, with a read N50 of 5.23 kb, corresponding to an estimated genome coverage of approximately 74.8–83.1× assuming a genome size of 450–500 Mb. For *Betta chloropharynx*, sequencing generated 14.97 million reads totaling 57.97 Gb, with a read N50 of 7.75 kb, representing an estimated coverage of approximately 115.9–128.8×. Meanwhile, *Betta schalleri* produced 11.83 million reads totaling 23.75 Gb, with a read N50 of 5.5 kb, corresponding to an estimated genome coverage of approximately 47.5–52.8×.

Genome assembly and annotation

The genome assembly and annotation described in this paper were conducted in accordance with the methodology performed by Imron et al.¹⁴ Unless otherwise stated, all software tools were run using their default parameters. Adapter contamination screening was performed using Porechop_ABI prior to assembly.¹⁵ Genome assembly was done using Flye 2.9.5,¹⁶ while genome scaffolding was conducted using RagTag 2.1.0¹⁷ guided by genome reference of *Betta splendens* (GCF_900634795.4). Foreign contamination was screened and removed using NCBI FCS-GX prior to scaffolding.¹⁸ Genome size estimation was conducted using Jellyfish software version 2.3.1¹⁹ and was further processed using GenomeScope 2.0 v2.0.1. The assembly statistics were calculated using assembly stat version 1.0.1. The completeness of the assembly was estimated using Benchmarking Universal Single-Copy Orthologous (BUSCO) version 5.8.2, utilizing miniprot^{20–22} with the Actinopterygii lineage dataset.

Repetitive elements within the genome assembly were identified using RepeatModeler v2.0.6 in conjunction with RepeatMasker v4.1.7. (<http://www.repeatmasker.org>). Prior to annotation, these repetitive regions were soft masked to minimize interference. Structural genome annotation encompassing gene prediction was conducted using the GALBA pipeline,²³ which employs miniprot²² and AUGUSTUS,²⁴ integrating protein data from closely related species as extrinsic evidence. Specifically, protein data from *Betta splendens* (GCF_900634795.4), *Anabas testudineus* (GCF_900324465.2), and *Channa argus* (GCF_033026475.1) were utilized. Functional annotation of the resulting gene predictions was then performed using the 'funannotate annotate' command from the Funannotate pipeline (<https://funannotate.readthedocs.io/en/latest/install.html>), incorporating tools such as InterProScan5,²⁵ EggNog-Mapper,²⁶ and SignalP 5.0²⁷ to assign gene names and predict protein functions. Finally, the completeness of the genome annotation was evaluated using BUSCO v5.8.2.²¹ with the Actinopterygii lineage dataset. For annotation-level assessment, BUSCO analysis was performed on the complete predicted protein dataset generated from the annotation pipeline, including all annotated transcript isoforms rather than restricting the analysis to only the longest isoform per gene.

Duplicate genes classification and comparative genomics

The protein sequences in each genomes were aligned to each other using BLAST+ v. 2.14.1 (<https://blast.ncbi.nlm.nih.gov/doc/blast-help/downloadblastdata.html>). The alignment results and the gff (general feature format) file containing gene position in the genome were then further analyzed using the 'Duplicate_gene_classifier' function of MCSanX²⁸ to identify gene duplications.

We then performed all-against-all alignment of protein sequences of three *Betta* genomes: using BLAST+ v. 2.14.1 (<https://blast.ncbi.nlm.nih.gov/doc/blast-help/downloadblastdata.html>). The analyses were performed using an E-value threshold of 1e−3, with results limited to a maximum of five target hits per query sequence and output generated in tabular format (outfmt 6). Computations were executed using 12 parallel threads, while all other parameters were retained at their default settings. The alignment results were then concatenated into one file. The gff files were also concatenated. These files were then further analyzed using MCSanX²⁸ to detect gene synteny and collinearity. The results were then explored and visualize using SynVisio.²⁹

We also performed comparative analysis of Colony Stimulating Factor 1 Receptor A (CSF1RA), Melanocortin 1 Receptor (MC1R) and Paired Box 7 (PAX7) protein sequence among *B. burdigala*, *B. chloropharynx*, and *B. schalleri*, alongside *B. splendens*, *Anabas testudineus*, *Channa argus*, *Oreochromis niloticus*, *Danio rerio*, and *Carassius auratus*. The protein accession of CSF1RA that we used are: XP026135817.1, XP026135818.1, NP571747.1, XP021336731.1, XP003455234.1, XP013133007.1, XP067373776.1, XP067373775.1, XP026213579.1, XP055368253.1, XP055368253.1, and XP029020690.1. The protein accession of MC1R that we used are: NP_851301.1, XP_005159236.1, XP_026112973.1, XP_026112974.1, XP_005467175.1, XP_055363428.1, XP_067343026.1, XP_067343026.1, XP_026234205.1. The protein accession of PAX7 that we used are: XP_025763203.1, XP_005459058.1, XP_067361857.1, XP_029005486.1, XP_029005485.1, XP_026204402.1, XP_026204405.1, XP_025763202.1, XP_067361858.1, XP_026204404.1, XP_029005487.1, XP_026204403.1, NP_571407.2, NP_571400.1, XP_009304561.1, XP_026130952.1, XP_026130953.1. The protein sequence alignments were performed using Clustal Omega³⁰ via the European Bioinformatics Institute website (<https://www.ebi.ac.uk/jdispatcher/msa/clustalo>). Subsequently, these alignments were used for phylogenetic reconstruction with IQ-TREE,³¹ and the results were visualized using TreeViewer.³²

Results

Fish morphology

The morphology photo of Betta spp. is presented in this research as shown in Figure 2 for *B. burdigala* (A), *B. chloropharynx* (B), and *B. schalleri* (C). Table 1 outlines the meristic characteristics that distinguish them. The meristic and morphometric characters are presented in Suppl Fig 1.

Betta burdigala

Betta burdigala is classified within the *Betta coccinea* group.³³ This species is characterized by its red coloration and a size range of 2-3 cm, with distinct morphological features illustrated in Figure 2A that differentiate it from other species within the *Betta coccinea* group. The meristic characteristics of *B. burdigala* are comprehensively presented in Table 1. In comparison to *B. uberis*,³⁴ *B. burdigala* exhibits fewer dorsal fin rays (14-15 vs. 14-17), subdorsal scales (11-11.5 vs. 12-13.5), predorsal scales (15-16 vs. 18-20), and postdorsal scales (8 vs. 9-11). Regarding morphometric characteristics, *B. burdigala* demonstrates a longer postdorsal length as a percentage of standard length (12.19-20.85% vs. 13.5-17.8%), a shorter dorsal fin base (28.12-32.03% vs. 30.0-37.2%), and a greater postorbital length (23.20-41.36% vs. 13.4-16.8%).

Betta chloropharynx

Betta chloropharynx, depicted in Figure 2B, is a member of the *Betta waseri* group.³⁵ This group is distinguished by a unique pattern on the ventral side of the head. *Betta chloropharynx* can be differentiated from *B. hipposideros* by several characteristics: it possesses ω -shaped black throat markings as opposed to horseshoe-shaped ones, lacks transverse lines on the caudal fin, and has fewer subdorsal scales, typically 5-6 compared to 6.5 (Table 1). In comparison to *B. renata*, *B. chloropharynx* exhibits ω -shaped black throat marks instead of kidney-shaped ones, lacks transverse lines on the dorsal and caudal regions, has an unspotted operculum rather than a speckled one, and features a yellow underside on the operculum instead of a black edge, with fewer subdorsal scales, mode 5-6 versus 6.5. The distinctions between

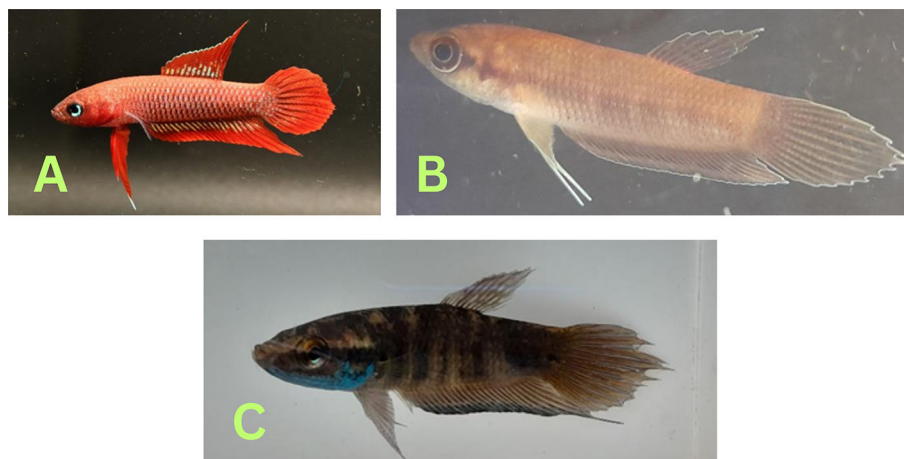


Figure 2. The morphology of *B. burdigala* (A), *B. chloropharynx* (B) and *B. schalleri* (C).

Table 1. Meristic and morphometric character of *B. burdigala*, *B. chloropharynx* and *B. schalleri*. The average of morphometric characters as a percentage to standard length with standard deviation.

	<i>B. burdigala</i>	<i>B. chloropharynx</i>	<i>B. schalleri</i>
Meristic character			
Dorsal fin radii	D.I. 13–14	D.I.8-9	D.I.8-9
Anal fin radii	A.I-II. 23–24	A.I-II. 27-28	A.I-II.22-23
Caudal radii	C.XV	C.XV	-
Standard length (cm)	2.46 ± 0.13	9.95 ± 0.91	5.21 ± 1.23
Morphometric character			
Post-orbital length (%)	34.71 ± 8.17	47.60 ± 1.06	45.52 ± 0.74
Dorsal fin length (%)	30.43 ± 1.68	14.00 ± 0.29	15.01 ± 0.95
Predorsal length (%)	51.96 ± 2.20	65.17 ± 0.81	63.01 ± 1.26
Postdorsal length (%)	17.33 ± 3.72	22.45 ± 1.69	21.16 ± 1.28
Orbital diameter (%)	6.52 ± 1.81	9.63 ± 0.51	9.24 ± 0.75
Body depth (%)	20.48 ± 1.28	25.78 ± 1.68	27.59 ± 0.25
Head length (%)	29.04 ± 0.60	33.97 ± 1.40	34.90 ± 1.90
Anal fin length (%)	46.15 ± 5.16	49.30 ± 0.17	47.73 ± 1.91
Preanal length (%)	47.73 ± 3.99	47.45 ± 1.16	47.91 ± 0.53

B. chloropharynx and *B. spilotogena* include ω-shaped black throat marks versus a central black spot, an unspotted operculum versus a spotted one, and an operculum with a yellow underside versus a posterior edge.³⁶

Betta schalleri

In [Figure 2C](#), *Betta schalleri*, when compared to other *B. pugnax*,³⁶ demonstrates distinct meristic characteristics. It possesses a greater number of anal fin rays (27 vs. 23–25), a higher count of dorsal fin rays than *B. cracens* and *B. fusca* (10–11 vs. 8–9), more subdorsal scales than *B. cracens* and *B. fusca* (6.5–7 vs. 5.5–6), a greater number of lateral scales than *B. fusca* (31 vs. 29), fewer lateral scales than *B. cracens* (31 vs. 32–33), and a reduced number of predorsal scales compared to *B. cracens*, *B. fusca*, and *B. raja* (17–19 vs. 20–24). In term of morphometric characteristics (expressed as a percentage of standard length; [Table 1](#)), *B. schalleri* exhibits a longer head length compared to *B. pugnax*, *B. cracens*, and *B. fusca* (35.5–36.5% vs. 27.5–35.2%), a shorter predorsal length compared to *B. fusca* (62.7–66.3% vs. 68.5–70.2%), a longer preanal length compared to *B. cracens* (47.16–48.30 % vs. 42.0–46.1%), a greater body depth compared to *B. cracens* (27.38–27.93 % vs. 21.2–24.2%), an extended dorsal fin base compared to *B. cracens* (14.11–16.32% vs. 10.5–11.6%), and a shorter anal fin base compared to *B. cracens* (45.49–50.16 % vs. 53.4–55.7%).

Genome assembly

The 21-mer profiling using GenomeScope indicated that the three endemic *Betta* species possess moderately compact genomes with relatively low heterozygosity levels, with *Betta schalleri* exhibiting the largest estimated genome size, followed by *Betta chloropharynx* and *Betta burdigala* ([Figure 3](#)). The genome assemblies of *B. burdigala*, *B. schalleri*, and *B. chloropharynx* demonstrated high contiguity and completeness, with *B. chloropharynx* showing the strongest contiguity, while *B. burdigala* and *B. schalleri* also provide high-quality genomic resources suitable for comparative and evolutionary studies ([Figure 3](#)). The genome assembly of *B. burdigala* totaled ~422 Mbp across 4,868 scaffolds and 6,904 contigs, with a scaffold N50 of 18 Mb and the longest scaffold reaching 33.2 Mb. The assembly was highly complete, with a BUSCO score of 99.1% (98.6% single-copy, 0.5% duplicated), while the GC content was 45.2% and gaps accounted for only 0.048%. The *B. chloropharynx* genome was slightly larger at ~474 Mbp, assembled into 3,273 scaffolds and 4,918 contigs. It showed the greatest contiguity, with a scaffold N50 of 20 Mbp, a maximum scaffold length of 35.9 Mbp, and the highest mean scaffold length (~145 kb). BUSCO completeness reached 99.3% (98.7% single-copy, 0.6% duplicated), the highest among the three species, with minimal fragmentation (0.5%) and missing genes (0.2%). The *Betta schalleri* assembly spanned ~433 Mbp with 8,474 scaffolds and 10,762 contigs, reflecting greater fragmentation. The scaffold N50 was 19 Mb, with the longest scaffold at 34.7 Mbp and a mean length of ~51 kb. BUSCO assessment indicated 99.1% completeness (98.8% single-copy, 0.4% duplicated), with GC content at 44.8% and low gap content (0.053%).

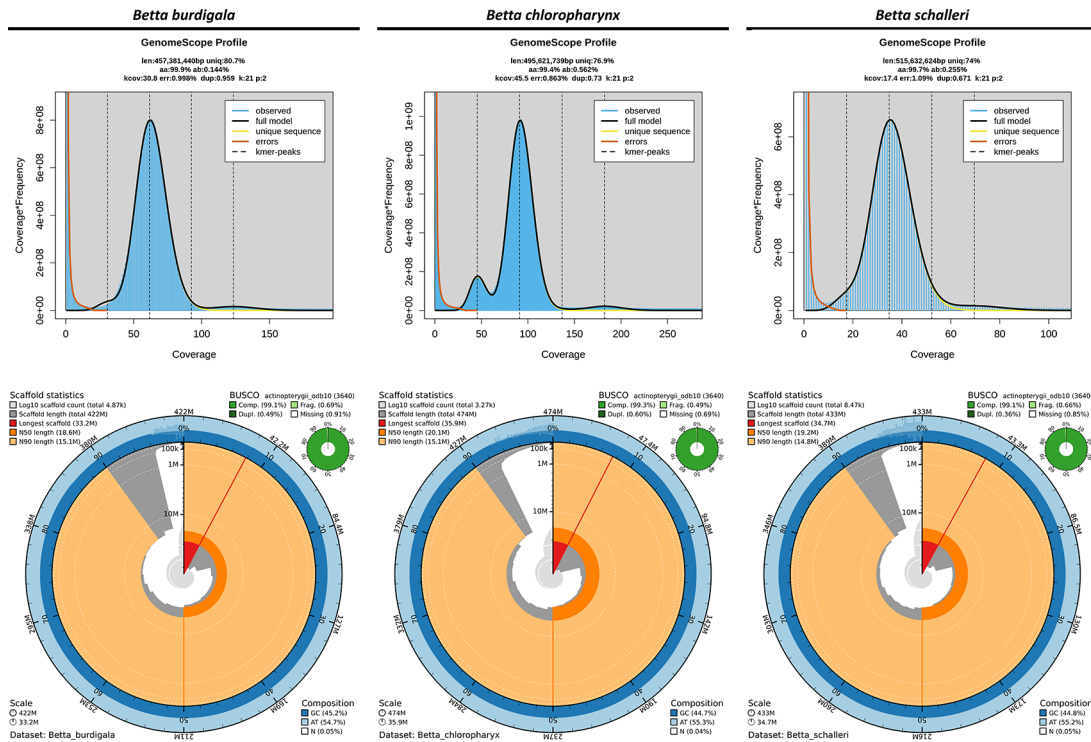


Figure 3. 21-mer-based genome size estimation (upper panels) and summary of genome assembly characteristics (lower panels) for *B. burdigala*, *B. chloropharynx*, and *B. schalleri*.

Genome annotation

Repeats annotation

Annotation of repetitive elements in the genomes of *B. burdigala*, *B. chloropharynx*, and *B. schalleri* revealed that transposable elements (TEs) comprise a substantial portion of each genome, with notable variation in both content and composition among species (Table 2). Retroelements were a dominant repeat category, accounting for 11.3% in BBS, 11.4% in *B. burdigala*, and up to 15.9% in *B. chloropharynx*. Among these, LINES were the most prevalent, particularly the L2/CR1/Rex and L1/CIN4 subfamilies, followed by LTR elements such as Gypsy/DIRS1 and BEL/Pao, which collectively contributed to the structural and regulatory diversification of the genome. Although SINES and Penelope elements were detected, they represented only a minor fraction of the total retroelement content (<0.3%).

DNA transposons were the second most abundant repeat class, comprising 6.5%, 9.8%, and 7.0% of the genomes of *B. burdigala*, *B. chloropharynx*, and *B. schalleri*, respectively (Table 2). The dominant families within this group included hobo-Activator, Tc1-IS630-Pogo, and En-Spm elements. These transposons are known for their cut-and-paste mechanism, and their relative abundance suggests that DNA-mediated transposition has played an important role in shaping genome architecture, particularly in BBS where their proportion was the highest. Interestingly, MULE-MuDR and PiggyBac elements were detected at low levels or were absent, suggesting potential lineage-specific losses or underrepresentation due to assembly or annotation limitations.

In addition to classified repeat categories, a significant fraction of the genome (3.9–5.9%) consisted of unclassified elements, indicating the presence of either novel transposable elements or highly diverged copies that defy current classification models (Table 2). Moreover, 22–30% of repetitive content across the three genomes fell into the “Other” category, representing repeat families that could not be assigned confidently to standard classifications. The higher repeat content observed in *B. chloropharynx*, particularly in retrotransposons and unclassified repeats, may reflect recent TE amplification events or reduced efficacy of TE suppression mechanisms in this lineage.

Structural and functional annotation and gene duplication analysis

The genome annotation summary across the three *Betta* species—*B. burdigala*, *B. chloropharynx*, and *B. schalleri*—revealed a broadly similar genomic architecture, with minor differences in gene structure metrics (Table 3). All three

Table 2. Classification of repeat elements of *B. burdigala*, *B. chloropharynx*, and *B. schalleri* genome assemblies.

Repeat category	Percentage of sequence (%)		
	<i>Betta burdigala</i>	<i>Betta chloropharynx</i>	<i>Betta schalleri</i>
Retroelements	5.98	8.39	6.18
SINEs	0.21	0.42	0.39
Penelope	0.03	0.01	0.32
LINEs	4.03	5.24	3.53
CRE/SLACS	0	0	0
L2/CR1/Rex	2.53	3.31	2.22
R1/LOA/Jockey	0.17	0.24	0.11
R2/R4/NeSL	0.25	0.37	0.36
RTE/Bov-B	0.41	0.8	0.54
L1/CIN4	0.58	0.35	0.16
BEL/Pao	0.11	0.25	0.21
Ty1/Copia	0.01	0.02	0.03
Gypsy/DIRS1	0.97	1.53	1.14
Retroviral	0.43	0.65	0.32
DNA transposons	3.85	5.54	4.02
hobo-Activator	1.18	1.86	1.5
Tc1-IS630-Pogo	0.68	1.65	1.08
En-Spm	0	0	0
MULE-MuDR	0	0	0
PiggyBac	0.13	0.2	0.1
Tourist/Harbinger	0.62	0.53	0.32
Rolling-circles	0.04	0.19	0.03
Unclassified	4.21	5.93	3.94
Total Interspersed	14.03	19.86	14.14
Small RNA	0.12	0.34	0.25
Satellites	0.06	0.07	0.05
Simple repeats	2.14	1.59	1.82
Low complexity	0.23	0.21	0.22
Total Masked Bases	16.54	22.08	16.30

species exhibit approximately 1.3 transcripts per gene on average, and the number of predicted transcripts ranged from 31,517 in *B. burdigala* to 32,786 in *B. chloropharynx*. The mean exon size was consistent (172–174 bp), while the mean gene locus size ranged from 8,784 bp in *B. burdigala* to 9,013 bp in *B. chloropharynx*. Notably, the percentage of genes with alternative transcript variants was relatively stable across species (~23%), and most genes were multi-exonic (over 91%), indicating conserved splicing complexity in the genus.

Regarding genome composition, the GC content of genes and introns was consistent across species at 46% and 44%, respectively (Figure 4A). Exonic GC content was slightly higher, ranging from 54% in *B. chloropharynx* and *B. schalleri* to 55% in *B. burdigala* (Figure 4B). Exons accounted for 9–10% of the genome, while genes (including exons and introns) occupied approximately 47–49% (Figure 4B). Introns alone contributed 38–40% of the genome, reflecting their substantial role in genome size and structure (Figure 4B). These values suggest that, despite minor quantitative differences, the overall organization and composition of the genomes are highly conserved across the *Betta* species analyzed.

Table 3. Genome annotation summary for the genome assembly of *Betta burdigala*, *Betta chloropharynx*, and *Betta schalleri*.

Statistic	<i>Betta burdigala</i> (BBB)	<i>Betta chloropharynx</i> (BBC)	<i>Betta schalleri</i> (BBS)
Max transcripts/gene	12	11	11
Mean exon size (bp)	173.2	174.1	172.7
Mean gene locus size (bp)	8784.6	9013.1	8942.5
Mean distinct exons/gene	10.4	10.2	10.4
Mean transcripts/gene	1.3	1.3	1.3
Mean transcript size (bp)	2013.8	1995.4	2016.1
Number of distinct exons	245812	251941	247092
Number of genes	23645	24597	23767
Genes with alt. transcripts	5,440 (23.0%)	5,615 (22.8%)	5,479 (23.1%)
Multi-exon genes	21,835 (92.3%)	22,592 (91.8%)	21,922 (92.2%)
Predicted transcripts	31517	32786	31694
Single-exon genes	1,810 (7.7%)	2,005 (8.2%)	1,845 (7.8%)

The BUSCO analysis using the *actinopterygii_odb10* dataset revealed high completeness across all three *Betta* species proteomes, indicating high-quality gene annotations (Figure 4C). *Betta burdigala* showed the highest completeness with 98.0% of BUSCOs identified as complete, closely followed by *B. schalleri* at 97.9%, and *B. chloropharynx* at 97.8%. The majority of these complete BUSCOs were single-copy, representing 78.5–78.8% of the total, while duplicated BUSCOs accounted for approximately 19.2–19.3%. This elevated duplication level likely reflects the inclusion of multiple

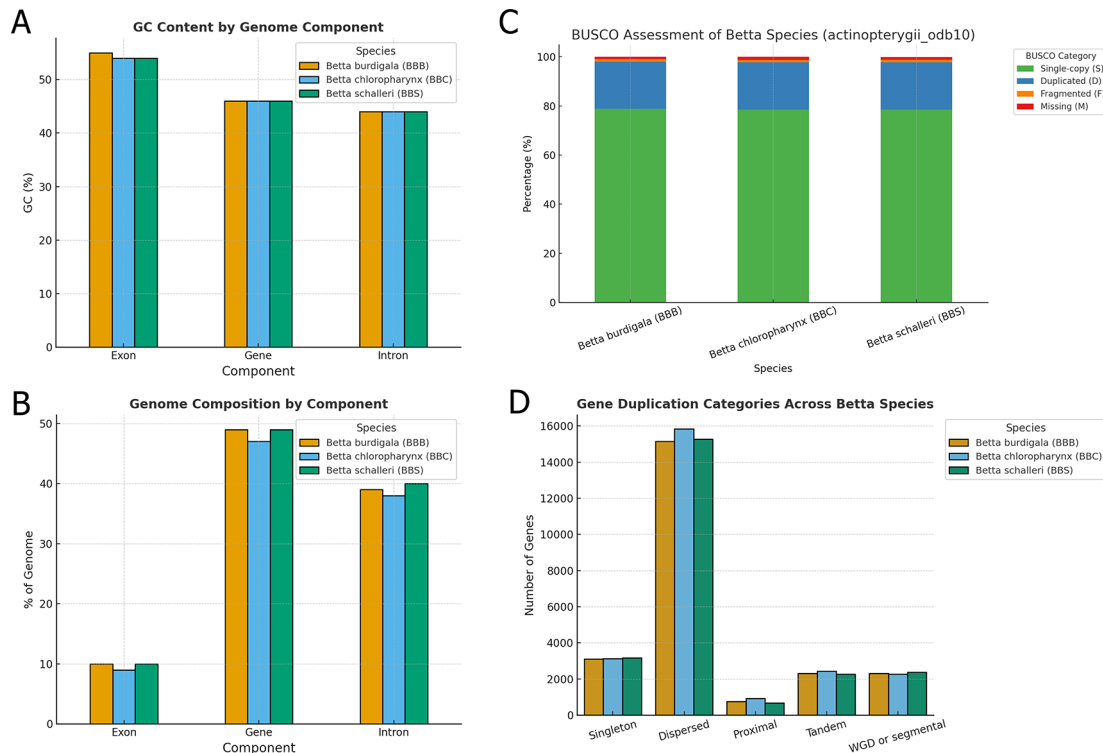


Figure 4. Genome annotation statistics and assessment of *Betta burdigala*, *Betta chloropharynx*, and *Betta schalleri*. (A) GC content by component of the three betta species genome assemblies. (B) Genome composition by component of the three Betta species genome assemblies. (C) BUSCO assessment of genome annotation results for the three Betta species genome assemblies. (D) Gene duplication classification in the genome of the three Betta species. Note that a gene can be classified into more than one gene duplication type.

transcript isoforms in the complete predicted protein sets rather than extensive true biological duplication. Nevertheless, the high proportion of complete BUSCOs supports the overall completeness of the genome annotations.

We also explored the gene duplication patterns revealing notable similarities and subtle differences in the distribution of duplication types among the three *Betta* species (Figure 4D). Across all species, dispersed duplications were the most prevalent, with *B. chloropharynx* showing the highest count (15,831 genes), followed closely by *B. schalleri* (15,269 genes) and *B. burdigala* (15,157 genes). This suggests that dispersed duplications represent a dominant mechanism contributing to gene expansion in these genomes, likely reflecting ongoing evolutionary pressures and functional diversification.

Other duplication categories, including tandem, proximal, and whole-genome duplication (WGD)/segmental events, were represented in roughly similar proportions across species (Figure 6). Tandem duplicates ranged from 2,277 in *B. schalleri* to 2,437 in *B. chloropharynx*, while proximal duplications were more variable, with *B. schalleri* showing the lowest count (673) compared to 924 in *B. chloropharynx*. WGD or segmental duplicates remained relatively consistent across species (approximately 2,270–2,370), highlighting a shared genomic history of large-scale duplication events. Singleton genes, which lack detectable paralogs, accounted for 3,114–3,178 genes depending on the species. The identified duplication patterns should therefore be interpreted as comparative genomic patterns inferred from predicted gene models and homologous gene relationships, rather than as definitive evidence of fully functional duplicated genes.

Comparative genomic

We conducted a collinearity analysis to investigate genome-level relationships among *B. burdigala*, *B. chloropharynx*, and *B. schalleri*. The analysis identified conserved syntenic gene blocks between each pair of species, which were subsequently quantified as collinear gene pairs (Figure 5A). A total of 16,948 pairs were detected between *B. burdigala* and *B. chloropharynx* (33.15%), 16,778 pairs between *B. burdigala* and *B. schalleri* (32.81%), and 17,405 pairs between *B. chloropharynx* and *B. schalleri* (34.04%). These values indicate a broadly conserved syntenic patterns across all three species, with each pair retaining roughly one-third of the total detected collinear relationships.

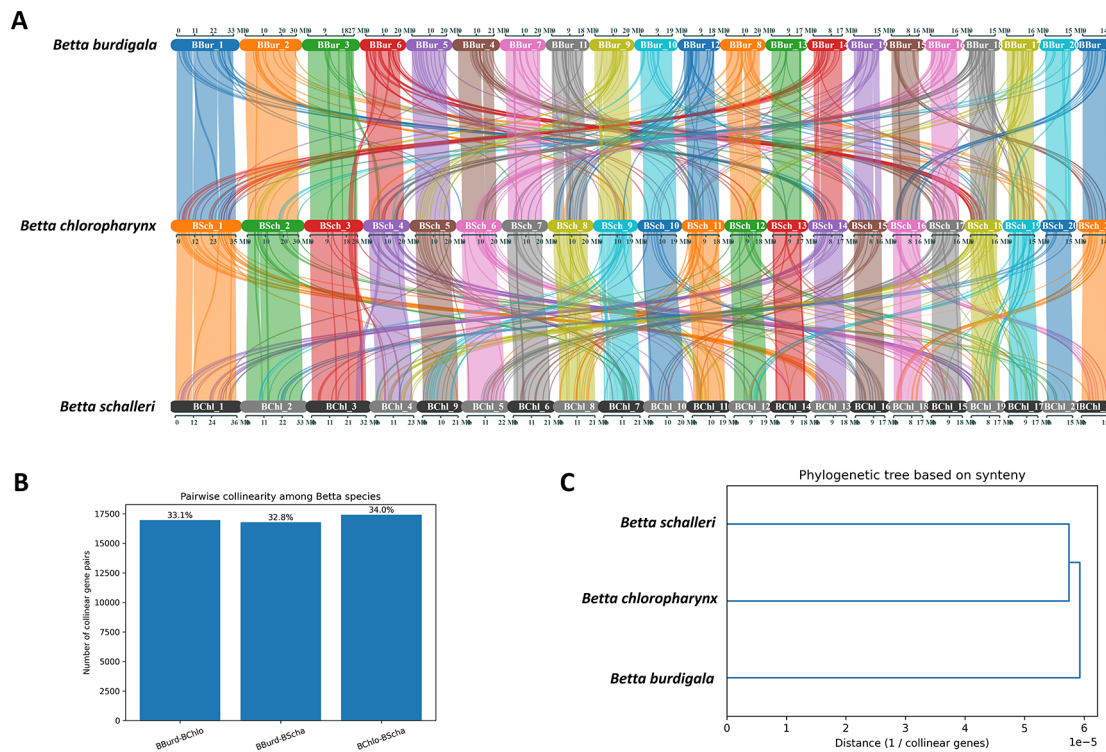


Figure 5. Collinearity analysis of genome assembly of *Betta burdigala* (BBur), *Betta schalleri* (BSch), and *Betta chloropharynx* (BChl). (A) syntenic block connection of the three genomes. (B) Pairwise collinearity among Betta species. (C) Pylogenetic tree based on synteny.

Among the three comparisons, the highest level of collinearity was observed between *B. chloropharynx* and *B. schalleri*, suggesting that these two species are slightly more closely related to each other than to *B. burdigala* (Figure 5B). In contrast, *B. burdigala* shows a similar but marginally lower level of collinearity with both *B. chloropharynx* and *B. schalleri*, implying that *B. burdigala* likely diverged earlier from the lineage leading to *B. chloropharynx* and *B. schalleri*. Visualization of the results using a dendrogram and a network graph further supports this interpretation. The dendrogram places *B. chloropharynx* and *B. schalleri* as sister species, with *B. burdigala* forming an outgroup (Figure 5C). The network graph illustrates conserved syntenic connections among all three species, but with the thickest edge linking *B. chloropharynx* and *B. schalleri*, reflecting their higher degree of shared genome organization.

We also did comparative analysis of three genes that might involve in color patterning of fishes which are Colony Stimulating Factor 1 Receptor A (CSF1RA), Melanocortin 1 Receptor (MC1R) and Paired Box 7 (PAX7). The protein sequence alignment are available in supplementary files (Supplementary file 1, Supplementary file 2, and Supplementary file 3). Gene-based phylogenetic analyses showed that the CSF1RA tree exhibited clustering patterns broadly consistent with the established taxonomic relationships among the nine fish species based on NCBI taxonomy (Figure 6A and D). We note that these analyses are based on single genes and therefore are intended primarily for comparative assessment of gene-level evolutionary patterns rather than for robust species-level phylogenetic inference. On the other hand, we observed that the branching pattern of the MC1R and PAX7 phylogenetic tree does not align with the species phylogenetic tree (Suppl Fig 2). We also identified two significant protein insertions in the CSF1RA of *B. burdigala* (Figure 6B and C). The first insertion, consisting of 27 amino acids, is located at approximately alignment position 594, just before the start of the Kinase Domain (GKVLGAGAFG...). The second insertion, comprising 28 amino acids, is situated at around alignment position 701, between the two lobes of the kinase domain.

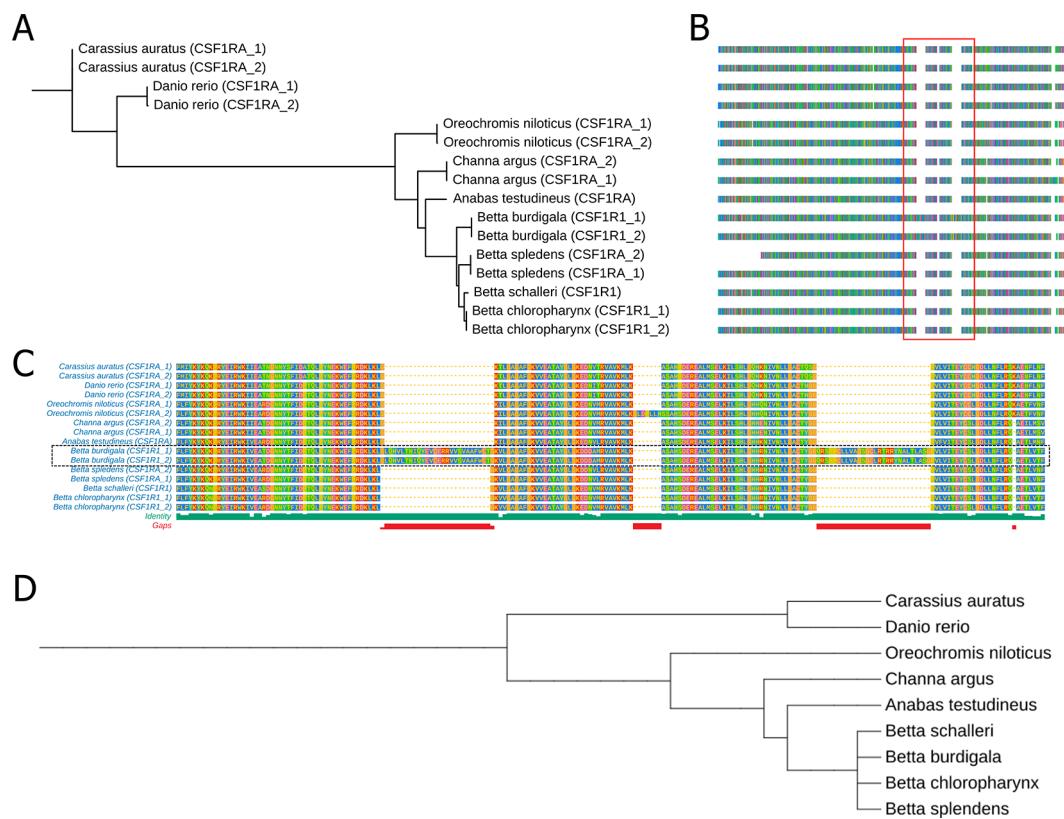


Figure 6. A comparative analysis of the Colony Stimulating Factor 1 Receptor A (CSF1RA) protein sequence among *B. burdigala*, *B. chloropharynx*, and *B. schalleri*, alongside *B. splendens*, *Anabas testudineus*, *Channa argus*, *Oreochromis niloticus*, *Danio rerio*, and *Carassius auratus*. (A) A phylogenetic tree derived from CSF1RA protein alignments across nine fish species. (B) Multiple Sequence Alignment of the CSF1RA protein in these nine fish species. (C) A zoomed-in view of the sequence alignment within the red line box in (B). (D) a phylogenetic tree of nine fish species based on NCBI taxonomic data. The complete protein sequence alignment of CSF1RA is available in Supplementary file 1.

Discussion

Our study integrates meristic analyses and genome assemblies to provide a comparative framework for understanding divergence among three endemic *Betta* species from Bangka Island: *B. burdigala*, *B. chloropharynx*, and *B. schalleri*. Morphological comparisons revealed clear diagnostic traits that distinguish these species and align with their current taxonomic placement. *B. burdigala* showed reductions in dorsal and predorsal scale counts, consistent with its placement in the *coccina* group.³³ *B. chloropharynx* was distinguished by its characteristic throat markings and opercular coloration typical of the *waseri* group,³⁵ while *B. schalleri* exhibited greater fin ray counts and deeper body proportions, traits associated with the *pugnax* group.³⁶ These meristic patterns reinforce species boundaries and highlight morphological divergence even within closely related taxa.

The genomic assemblies presented here represent the first references for these Bangka endemics. All three species demonstrated high contiguity and completeness, confirming their reliability for downstream analyses. Subtle genomic differences were also evident. *B. chloropharynx* displayed the largest genome size, highest scaffold N50, and elevated retrotransposon content, suggesting lineage-specific repeat dynamics. Gene duplication profiles showed dispersed duplications as the dominant category across all species, while variation in tandem and proximal duplications may indicate lineage-specific expansions.^{37–39}

These findings demonstrate that while overall genome organization and gene content remain conserved across the three species, lineage-specific signatures of divergence are detectable. The combination of high collinearity and subtle differences in repeat content and duplication patterns provides a genomic basis for understanding evolutionary relationships within the genus.^{40,41} Moreover, the integration of meristic and genomic evidence strengthens taxonomic resolution, particularly in lineages where morphology alone may be insufficient.

Together, the annotation and composition data indicate a strong conservation of gene architecture and genomic content within the *Betta* species, consistent with their relatively recent divergence. Subtle differences, such as the slightly higher transcript count or gene locus size in *B. chloropharynx*, may hint lineage-specific regulatory or structural variations.^{42,43} The consistent duplication profiles across the three species suggest conserved duplication dynamics within the *Betta* genus. However, the modest variation in tandem and proximal duplication counts may point to lineage-specific gene family expansions or contraction events, potentially linked to evolutionary divergence among these lineages.^{38,44} However, population-level genomic data will be required to definitively connect these structural variations to specific ecological adaptations.

Collinearity analysis revealed broadly conserved syntenic patterns among the three *Betta* species, with *B. chloropharynx* and *B. schalleri* showing comparatively higher levels of shared gene order and synteny relative to *B. burdigala*. These patterns are generally consistent with their closer evolutionary relationship inferred from existing taxonomic and comparative genomic evidence. Overall, the results suggest that the three *Betta* species retain largely conserved genome organization, while *B. chloropharynx* and *B. schalleri* may share more similar genomic architectures compared to *B. burdigala*. Because the assemblies were scaffolded using a reference-guided approach, these collinearity patterns should be interpreted cautiously as indicators of broad-scale syntenic conservation rather than definitive evidence of fully conserved chromosomal structure. Nevertheless, the analyses provide a useful comparative genomic framework for exploring genome organization and evolutionary relationships within the *Betta* species.^{42,45,46} The observed syntenic conservation, particularly between *B. chloropharynx* and *B. schalleri*, is also consistent with patterns reported in other closely related fish taxa, where substantial conservation of gene order is often retained despite lineage-specific chromosomal rearrangements in more divergent groups.^{47,48}

Our comparative genomic analysis reveals a two-layered evolutionary dynamic underlying pigmentation diversity in *Betta* species: strong conservation of core pigment-cell survival pathways alongside rapid divergence in genes governing color type and pattern formation. CSF1RA, a key regulator of melanophore and color pigments (xanthophores, erythrophores, and iridophore) development,^{49–51} exhibited clustering patterns broadly consistent with the accepted species relationships and shows high sequence conservation across all examined fishes, reflecting high sequence conservation and potential evolutionary constraints on this essential signaling pathway. Notably, however, *B. burdigala* exhibits two large amino-acid insertions positioned adjacent to and within the kinase domain, indicating species-specific modulation rather than disruption of receptor function. While further research is necessary, these insertions may contribute to modulation of pigment patterning or coloration intensity, potentially enhancing this species' vibrant coloration. In contrast, genes associated with pigment-cell fate and color intensity—particularly PAX7 and MC1R—showed topologies that differed from the accepted species relationships. Divergence in PAX7, a major regulator of xanthophore and erythrophore lineages,^{52–54} is consistent with adaptive shifts affecting red–yellow pigmentation, especially relevant given the pronounced red coloration in *B. burdigala*. Similarly, the non-concordant evolution of MC1R, despite its variable role in teleost pigmentation,^{55–57} is consistent with high sequence divergence that may

warrant future investigation into potential regulatory or functional divergence within the *Betta* genus. Together, these findings indicate that *Betta* color diversity arises from conserved developmental frameworks overlaid by lineage-specific sequence variation and structural innovation, enabling fine-scale modulation of pigment cell types and color intensity across species.⁵⁸

Although our data do not directly address ecological or paleogeographic processes, previous studies suggest that the high endemism of Bangka's *Betta* species has been shaped by Pleistocene paleogeography and the persistence of specialized peat swamp habitats.^{6,59–64} In this context, the divergence patterns we observed in both morphology and genomics are consistent with scenarios of historical geographic isolation and ecological filtering proposed by earlier work. Our results therefore provide the genetic and morphological evidence that complements these broader biogeographic hypotheses. We acknowledge that the genome assemblies generated in this study are based on single individuals for each species and therefore may not fully represent the extent of intraspecific genomic variation present within natural populations. Nevertheless, these genomic resources contribute important data for species delimitation and phylogenomic placement within the *Betta* genus, while also establishing a foundation for future studies investigating the relationships among genomic variation, ecology, adaptation, and conservation. Given the conservation status of these taxa, the genomic resources generated here are valuable for monitoring genetic diversity and informing management strategies for Bangka's unique freshwater biodiversity.

Conclusion

This study provides the first integrated meristic and genomic comparison of three endemic *Betta* species from Bangka Island—*B. burdigala*, *B. chloropharynx*, and *B. schalleri*. Meristic analyses confirmed diagnostic traits supporting their taxonomic distinction, while high-quality genome assemblies with near-complete BUSCO scores (>97%) offer reliable references for evolutionary and conservation studies. Although overall genomic architecture is conserved, subtle differences in genome size, repeat composition, and duplication profiles highlight lineage-specific divergence, with synteny analyses indicating a closer relationship between *B. chloropharynx* and *B. schalleri*. By combining morphological and genomic evidence, this study not only strengthens the resolution of species boundaries but also establishes essential resources for monitoring genetic diversity and informing conservation strategies for these endangered freshwater fishes.

Data availability

The data related to this project have been deposited under NCBI BioProject Accession PRJNA1328177 and are publicly available. The biosample accession numbers are SAMN51307167, SAMN51307168, and SAMN51307169 for *B. burdigala*, *B. schalleri*, and *B. chloropharynx* respectively. The raw genome sequence reads have been deposited in the NCBI Sequence Read Archive (SRA) under the accession numbers SRR35652347 for *B. burdigala*, SRR35653255 for *B. schalleri*, and SRR35728893 and SRR35728892 for *B. chloropharynx*. Additionally, the genome assemblies have been deposited in the NCBI GenBank database with accession numbers GCA_054471155.1, GCA_054471145.1, GCA_054471135.1 for *B. burdigala*, *B. schalleri*, and *B. chloropharynx*, respectively.

The genome assemblies and their annotations have also been deposited and are available in the National Scientific Repository of the National Research and Innovation Agency (Badan Riset dan Inovasi Nasional/BRIN) of the Republic of Indonesia: Genome *Betta* spp. (<https://hdl.handle.net/20.500.12690/RIN/LEOQPP>).⁶⁵

Extended data

All extended data have been deposited in figshare (DOI: <https://doi.org/10.6084/m9.figshare.31189546.v3>).⁶⁶ The extended data include the following:

Suppl Fig 1. The meristic (A) and morphometric (B) characters of *Betta* spp.

Suppl Fig 2. A comparative analysis of the Melanocortin 1 Receptor (MC1R) and Paired box 7 (PAX7) protein sequences among *B. burdigala*, *B. chloropharynx*, and *B. schalleri*, alongside *B. splendens*, *Anabas testudineus*, *Channa argus*, *Oreochromis niloticus*, *Danio rerio*, and *Carassius auratus*. (A) A phylogenetic tree derived from MC1R protein alignments across nine fish species. (B) Multiple Sequence Alignment of the PAX7 MC1R protein in these nine fish species. (C) A phylogenetic tree derived from PAX7 protein alignments across nine fish species. (D) Multiple Sequence Alignment of the PAX7 protein in these nine fish species. The complete protein sequence alignment of MC1R and PAX7 are available in Supplementary file 2 and Supplementary file 3 respectively.

Suppl Fig 3. Validation of the CSF1R1 insertion using Oxford Nanopore long-read alignments. Coverage depth across the insertion locus remained consistent (~70–88×), and multiple long reads spanned the insertion region continuously, supporting the authenticity of the predicted insertion and excluding local assembly artifacts.

Supplementary file 1. Protein alignment of CSF1R1 in 9 fishes.clustal_num

Supplementary file 2. Protein alignment of MC1R in 9 fishes.clustal_num

Supplementary file 3. Protein alignment of PAX7 in 9 fishes.clustal_num

Supplementary file 4. Coverage depth of CSF1R1_1 region by Oxford Nanopore long-read alignments.

Supplementary file 5. QUAST assembly quality assessment statistics for the three *Betta* genome assemblies.

Reporting guidelines

The complete checklists of ARRIVE 2.0 reporting guidelines is available at <https://doi.org/10.6084/m9.figshare.31136455>.¹⁰

All the underlying and extended data of this study are openly available under the terms of The [CC BY 4.0 license](#) (Creative Commons Attribution 4.0 International).

Acknowledgements

We extend our sincere appreciation to the Directorate of Research and Community Service (DPPM) of Ministry of Higher Education, Science and Technology, and the Chairman of LLDIKTI II for their financial support of this research. We also express our gratitude to the Rector of Universitas Muhammadiyah Palembang, the Chairman of LPPM Universitas Muhammadiyah Palembang, and the Dean of the Faculty of Agriculture at Universitas Muhammadiyah Palembang for their support in the execution of this research. Additionally, we acknowledge PT. Timah Tbk. and Yayasan Ikan Endemik Bangka Belitung for facilitating this research. The authors are grateful for the facilities and scientific and technical support provided by the Whole Genome Sequencing Laboratory, National Research and Innovation Agency through E-Layanan Sains, Badan Riset dan Inovasi Nasional. We also wish to thank Integrated Genome Factory (IGF) Faculty of Biology UGM for granting access to computing facilities. During the preparation of this work, we used Paperpal to improve grammar and sentence structure in the manuscript. After using this tool, we reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

1. Helmuzuryani H, et al.: **Exploring freshwater fish biodiversity using eDNA metabarcoding and traditional sampling to assess floodplain waters.** *Journal of Water and Land Development*. 2025; 100–112.
[Publisher Full Text](#)
2. Nur FM, et al.: **Diversity, distribution, and conservation status of Betta fish (Teleostei: Osphronemidae) in Aceh waters, Indonesia.** *Eur. Zool. J.* 2022; **89**: 142–151.
3. Tan HH: **Redescription of *Betta anabatoides* Bleeker, and a new species of *Betta* from West Kalimantan, Borneo (Teleostei: Osphronemidae).** *Zootaxa*. 2009; **2165**: 59–68.
4. Pammanasut P, et al.: **Discovery of wild populations of *Betta smaragdina* Ladiges, 1972 (Teleostei, Osphronemidae) in a western province of Thailand.** *Check List*. 2018; **14**: 1077–1082.
5. Wattanadilokchatkun P, et al.: **Integrative approach for landscape demography analysis of Plakad-Pa Pak-Tawan-Ok (*Betta siamorientalis*): Deciphering genetic and environmental factors in Eastern Thailand's conservation efforts.** *Glob. Ecol. Conserv.* 2024; **51**: e02870.
[Publisher Full Text](#)
6. Valen FS, et al.: **Unveiling the close relationship between *Betta burdigala* and *Betta uberis* through DNA barcoding based on COI Gene.** *IOP Conf. Ser. Earth Environ. Sci.* 2023; **1267**: 012066.
7. Rhie A, et al.: **Towards complete and error-free genome assemblies of all vertebrate species.** *Nature*. 2021; **592**: 737–746.
[PubMed Abstract](#) | [Publisher Full Text](#)
8. Huang Y, et al.: **Reference Genome Assembly of the Big Berry Manzanita (*Arctostaphylos glauca*).** *J. Hered.* 2022; **113**: 188–196.
[PubMed Abstract](#) | [Publisher Full Text](#)
9. Theissinger K, et al.: **How genomics can help biodiversity conservation.** *Trends Genet.* 2023; **39**: 545–559.
[PubMed Abstract](#) | [Publisher Full Text](#)
10. Helmuzuryani H, et al.: **Author Checklist of The ARRIVE guidelines 2.0 for Beyond the Beauty: Meristic and Genomic Signatures of Bangka's Endemic *Betta* Fishes.** *figshare*. 2026.
[Publisher Full Text](#)
11. Batubara JP, Laila K, Rumodang R, et al.: **The Effects of Different Dosages of Pasak Bumi (*Eurycoma longifolia* Jack) towards the Sex of *Betta* Fish Embryo.** *IOP Conf. Ser. Earth Environ. Sci.* 2022; **1118**: 012005.
12. Pattiasina BJ, Pattinasarany MM, Manuputty MMD, et al.: **Masculinization of beta fish larvae *Betta splendens* through the different treatment immersion of honey solution and larval age.** *IOP Conf. Ser. Earth Environ. Sci.* 2021; **797**: 012016.
13. American Veterinary Medical Association: **AVMA Guidelines for the euthanasia of animals.** 2020. Vol. Accessed 10 March 2021.
14. Imron I, et al.: **Draft genome assembly of the slender walking catfish, *Prophagorus nieuhofii* [version 1; peer review: 2 approved with reservations, 1 not approved].** *F1000Res*. 2025; **14**.
[Publisher Full Text](#)
15. Bonenfant Q, Noé L, Touzet H: **Porechop_ABI: discovering unknown adapters in Oxford Nanopore Technology sequencing reads for downstream trimming.** *Bioinform. Adv.* 2023; **3**: vbac085.
[Publisher Full Text](#)
16. Kolmogorov M, Yuan J, Lin Y, et al.: **Assembly of long, error-prone reads using repeat graphs.** *Nat. Biotechnol.* 2019; **37**: 540–546.
[PubMed Abstract](#) | [Publisher Full Text](#)

17. Alonge M, *et al.*: **Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing.** *Genome Biol.* 2022; **23**: 258.
[Publisher Full Text](#)
18. Astashyn A, *et al.*: **Rapid and sensitive detection of genome contamination at scale with FCS-GX.** *Genome Biol.* 2024; **25**: 60.
[Publisher Full Text](#)
19. Marçais G, Kingsford C: **A fast, lock-free approach for efficient parallel counting of occurrences of k-mers.** *Bioinformatics.* 2011; **27**: 764–770.
[PubMed Abstract](#) | [Publisher Full Text](#)
20. Simão F, Waterhouse R, Ioannidis P, *et al.*: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.** *Bioinformatics.* 2015; **31**: 3210–3212.
[Publisher Full Text](#)
21. Manni M, Berkeley MR, Seppey M, *et al.*: **BUSCO: Assessing Genomic Data Quality and Beyond.** *Curr. Protoc.* 2021; **1**: e323.
[Publisher Full Text](#)
22. Li H: **Protein-to-genome alignment with miniprot.** *Bioinformatics.* 2023; **39**.
[Publisher Full Text](#)
23. Brůna T, *et al.*: **Galba: genome annotation with miniprot and AUGUSTUS.** *BMC Bioinformatics.* 2023; **24**: 327.
[Publisher Full Text](#)
24. Hoff KJ, Stanke M: **Predicting Genes in Single Genomes with AUGUSTUS.** *Curr. Protoc. Bioinformatics.* 2019; **65**: e57.
[Publisher Full Text](#)
25. Jones P, *et al.*: **InterProScan 5: genome-scale protein function classification.** *Bioinformatics.* 2014; **30**: 1236–1240.
[PubMed Abstract](#) | [Publisher Full Text](#)
26. Cantalapiedra CP, Hernández-Plaza A, Letunic I, *et al.*: **eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale.** *Mol. Biol. Evol.* 2021; **38**: 5825–5829.
[PubMed Abstract](#) | [Publisher Full Text](#)
27. Almagro Armenteros JJ, *et al.*: **SignalP 5.0 improves signal peptide predictions using deep neural networks.** *Nat. Biotechnol.* 2019; **37**: 420–423.
[PubMed Abstract](#) | [Publisher Full Text](#)
28. Wang Y, *et al.*: **MCSANX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity.** *Nucleic Acids Res.* 2012; **40**: e49.
29. Bandi V, *et al.*: **Visualization Tools for Genomic Conservation.** *Methods Mol. Biol.* 2022; **2443**: 285–308.
[PubMed Abstract](#) | [Publisher Full Text](#)
30. Sievers F, Higgins DG: **Clustal Omega, accurate alignment of very large numbers of sequences.** *Methods Mol. Biol.* 2014; **1079**: 105–116.
[PubMed Abstract](#) | [Publisher Full Text](#)
31. Thomas KF, Wong NL-T, Ren H, *et al.*: **IQ-TREE 3: Phylogenomic Inference Software using Complex Evolutionary Models.** 2025.
[Reference Source](#)
32. Bianchini G, Sánchez-Baracaldo P: **TreeViewer: Flexible, modular software to visualise and manipulate phylogenetic trees.** *Ecol. Evol.* 2024; **14**: e10873.
33. Ding J, Lei W, Haryono H, *et al.*: **Phylogenetic analysis of *Betta coccina* complex (Teleostei, Osphronemidae) from Peninsular Malaysia and Sumatra Island with descriptions of two new species.** *ZooKeys.* 2025; **1238**: 161–181.
[PubMed Abstract](#) | [Publisher Full Text](#)
34. Tan HH, Ng PKL: **Six new species of fighting fish (Teleostei: Osphronemidae: *Betta*) from Borneo.** *Ichthyol. Explor. Freshwat.* 2006; **17**: 97–114.
35. Tan HH, Ahmad AB: ***Betta omega*, a new species of black water fighting fish (Teleostei: Osphronemidae) from Malaysia.** *Raffles Bull. Zool.* 2018; **66**: 402–407.
36. Tan HH, Ng PKL: **The labyrinth fishes (Teleostei: Anabantoidei, Channoidei) of Sumatra, Indonesia.** *Raffles Bull. Zool.* 2005; **13**: 115–138.
37. Chen Z, *et al.*: **De novo assembly of the goldfish (*Carassius auratus*) genome and the evolution of genes after whole-genome duplication.** *Sci. Adv.* 2019; **5**: eaav0547.
38. Lu J, Peatman E, Tang H, *et al.*: **Profiling of gene duplication patterns of sequenced teleost genomes: evidence for rapid lineage-specific genome expansion mediated by recent tandem duplications.** *BMC Genomics.* 2012; **13**: 246.
[Publisher Full Text](#)
39. Nakamura Y, *et al.*: **Rhodopsin gene copies in Japanese eel originated in a teleost-specific genome duplication.** *Zool. Lett.* 2017; **3**: 18.
[Publisher Full Text](#)
40. Li S, *et al.*: **Genome collinearity analysis illuminates the evolution of donkey chromosome 1 and horse chromosome 5 in perissodactyls: A comparative study.** *BMC Genomics.* 2021; **22**: 665.
[Publisher Full Text](#)
41. Wang X, *et al.*: **Evolutionary Divergence between *Toona ciliata* and *Toona sinensis* Assayed with Their Whole Genome Sequences.** *Genes.* 2022; **13**.
[Publisher Full Text](#)
42. Kwon YM, *et al.*: **Genomic consequences of domestication of the Siamese fighting fish.** *Sci. Adv.* 2022; **8**: eabm4950.
43. Zhang W, *et al.*: **The genetic architecture of phenotypic diversity in the Betta fish (*Betta splendens*).** *Sci. Adv.* 2022; **8**: eabm4955.
44. Huang P, *et al.*: **Chromosome-level genome assembly and population genetic analysis of a near-threatened rosewood species (*Dalbergia cultrata* Pierre Graham ex Benth) provide insights into its evolutionary and cold stress responses.** *Front. Plant Sci.* 2023; **14**.
[Publisher Full Text](#)
45. Fan G, *et al.*: **Chromosome-level reference genome of the Siamese fighting fish *Betta splendens*, a model species for the study of aggression.** *GigaScience.* 2018; **7**: gyl087.
[Publisher Full Text](#)
46. Srikulnath K, *et al.*: **Overview of the *Betta* fish genome regarding species radiation, parental care, behavioral aggression, and pigmentation model relevant to humans.** *Genes & Genomics.* 2021; **43**: 91–104.
[PubMed Abstract](#) | [Publisher Full Text](#)
47. Brasó-Vives M, *et al.*: **Parallel evolution of amphioxus and vertebrate small-scale gene duplications.** *Genome Biol.* 2022; **23**: 243.
48. Shen Q, *et al.*: **Chromosome-level genome assembly of the butterfly hillstream loach *Beaufortia pingi*.** *Sci. Data.* 2024; **11**: 1260.
[Publisher Full Text](#)
49. Patterson LB, Bain EJ, Parichy DM: **Pigment cell interactions and differential xanthophore recruitment underlying zebrafish stripe reiteration and Danio pattern evolution.** *Nat. Commun.* 2014; **5**: 5299.
50. Braasch I, Salzburger W, Meyer A: **Asymmetric Evolution in Two Fish-Specifically Duplicated Receptor Tyrosine Kinase Paralogons Involved in Teleost Coloration.** *Mol. Biol. Evol.* 2006; **23**: 1192–1202.
[PubMed Abstract](#)
51. Stanley ER, *et al.*: **Biology and action of colony-stimulating factor-1.** *Mol. Reprod. Dev.* 1997; **46**: 4–10.
[PubMed Abstract](#)
52. Nord H, Dennhag N, Muck J, *et al.*: **Pax7 is required for establishment of the xanthophore lineage in zebrafish embryos.** *Mol. Biol. Cell.* 2016; **27**: 1853–1862.
[PubMed Abstract](#) | [Publisher Full Text](#)
53. Roberts RB, Moore EC, Kocher TD: **An allelic series at pax7a is associated with colour polymorphism diversity in Lake Malawi cichlid fish.** *Mol. Ecol.* 2017; **26**: 2625–2639.
[PubMed Abstract](#) | [Publisher Full Text](#)
54. Pokrajac NT, Singh AP: **Spotting the band.** *Proc. Natl. Acad. Sci.* 2024; **121**: e2412259121.
55. Tezuka A, Yamamoto H, Yokoyama J, *et al.*: **The MC1R gene in the guppy (*Poecilia reticulata*): Genotypic and phenotypic polymorphisms.** *BMC. Res. Notes.* 2011; **4**: 31.
[Publisher Full Text](#)
56. Cal L, Suarez-Bregua P, Cerdá-Reverter JM, *et al.*: **Fish pigmentation and the melanocortin system.** *Comp. Biochem. Physiol. A Mol. Integr. Physiol.* 2017; **211**: 26–33.
[PubMed Abstract](#) | [Publisher Full Text](#)
57. Mundy NI: **Conservation and convergence of colour genetics: MC1R mutations in brown cavefish.** *PLoS Genet.* 2009; **5**: e1000388.
58. Maan ME, Sefc KM: **Colour variation in cichlid fish: Developmental mechanisms, selective pressures and evolutionary consequences.** *Semin. Cell Dev. Biol.* 2013; **24**: 516–528.
[PubMed Abstract](#) | [Publisher Full Text](#)
59. Ng PKL, Tan HH: **Freshwater fishes of Southeast Asia: potential for the aquarium fish trade and conservation issues.** *Aquar. Sci. Conserv.* 1997; **1**: 79–90.
60. Voris HK: **Maps of Pleistocene sea levels in Southeast Asia: shorelines, river systems and time durations.** *J. Biogeogr.* 2000; **27**: 1153–1167.
61. Kottelat M: **The fishes of the inland waters of Southeast Asia: a catalogue and core bibliography of the fishes known to occur in freshwaters, mangroves and estuaries.** *Raffles Bull. Zool.* 2013; **27**: 1–663.
62. Sathiamurthy E, Voris H: **Maps of Holocene Sea Level Transgression and Submerged Lakes on the Sunda Shelf.**

Tropical Nat. Hist. 2006; 1–44.

[Publisher Full Text](#)

63. Ramadhanu D, Prananda M, Wulandari UA, *et al.*: **Perbandingan Pola Rasio Morfometrik dan Karakteristik Habitat Dua Spesies Ikan Cupang Alam (*Betta spp.*) Endemik Pulau Bangka.** *Bioscientist: Jurnal Ilmiah Biologi.* 2023; **11**: 1501–1508.
64. Khanati O, Lista D, Lindiatika L, *et al.*: **Bioecology of Wild *Betta* Fish (*Betta schalleri*) of Bangka Island Waters.** *Jurnal Lahan*

Suboptimal: Journal of Suboptimal Lands. 2024; **13**.

[Publisher Full Text](#)

65. Helmizuryani H, *et al.*: **Genome *Betta spp.* V2 edn (RIN Dataverse).** 2025.
66. Helmizuryani H, *et al.*: **[Extended data] Beyond the Beauty: Meristic and Genomic Signatures of Bangka's Endemic *Betta* Fishes.** 2026.

[Publisher Full Text](#)

Open Peer Review

Current Peer Review Status:   

Version 2

Reviewer Report 13 June 2026

<https://doi.org/10.5256/f1000research.202000.r489884>

© 2026 Xing B. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Bingpeng Xing

Third Institute of Oceanography Ministry of Natural Resources, Xiamen, China

The authors have adequately addressed the concerns raised during the previous round of review. In particular, they have improved methodological transparency, added genome quality metrics and contamination screening, clarified the limitations of reference-guided scaffolding and single-gene phylogenetic analyses, and provided additional evidence supporting the reported CSF1RA insertions. The remaining issues are mainly minor language and copyediting problems and do not affect the validity of the data or the conclusions. I believe the manuscript is now suitable for indexing and recommend approval.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: marine biology, genome, fish, genetic, DNA, eDNA

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 10 June 2026

<https://doi.org/10.5256/f1000research.202000.r491035>

© 2026 Janke A. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Axel Janke

Senckenberg Biodiversity and Climate Research Centre, Frankfurt am Main, Germany

The genome data are interesting and clearly worth publishing. The authors present a high-quality assembly of yet another species in the genus, supported by standard descriptive analyses, clear

graphics, and some intra-genus comparisons. Overall, the technical work is solid and well executed.

However, I find the biological and scientific novelty rather limited. While the genome itself is a useful addition, I am not particularly convinced by the single-gene phylogenetic analyses presented, especially given that the full genome is available. More comprehensive genome-wide analyses would have strengthened the study significantly.

If possible, cite the first Betta genome: (Ref 1 - Prost S, et al., 2020) Prost S, Petersen M, Grethlein M, et al. Improving the Chromosome-Level Genome Assembly of the Siamese Fighting Fish (*Betta splendens*) in a University Master's Course. G3 (Bethesda, Md.). 2020 Jul;10(7):2179-2183. DOI: 10.1534/g3.120.401205. PMID: 32385046; PMCID: PMC7341155.

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Mammalian evolution, population genomics, giraffe, bear, whales, wolves, wolverines, genome assembly

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 05 March 2026

<https://doi.org/10.5256/f1000research.192703.r457789>

© 2026 Kron N. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**Nicholas S Kron** ¹ University of Miami, Rickenbacker Cswy, Miami, USA² Genevia Technologies, Finland, Finland

Summary

In "Beyond the Beauty: Meristic and Genomic Signatures of Bangka's Endemic Betta Fishes", Helmizuryani et al. describe the collection, sampling, genome assembly and annotation, and comparative genomic analysis of three species of the genus *Betta* from the Bangka islands of Indonesia. The authors demonstrate high quality, sub-chromosome level genome assemblies using Oxford Nanopore sequencing and established long read assembly and annotation workflows. Based on genome annotation results, the authors describe the genomes of the three species as highly colinear and consistent, with minor variation among them possibly relating to evolutionary divergence among the three, with special attention paid to pigmentation associated genes.

Comments

The study employs commonly used open-source tools and standard methodologies for assembly and annotation. The genomic resources generated by this study are novel and valuable contributions and will surely be useful for future studies in these species and for teleost and vertebrate genomics more broadly.

The manuscript would benefit from proof reading for spelling, grammar, consistency, etc.:

- Consistency of italicization for species names throughout.
- Introduction: "The number of endemic wild Betta distribution and diversity of Southeast Asia has been identified for 16 years in Thailand (2 species) and Indonesia (9 species)"
- Introduction: "Three species of Betta spp. were endemic, i.e., *Betta burdigala*, *Betta chloropharynx*, and *Betta schalleri*."
- Introduction: "The genetic studies of the endemic fish Betta spp. on Bangka Island have..."
- Introduction: "However, the whole genome for the evolution of the endemic fish Betta spp. on Bangka Island hasn't been revealed."
- Introduction: "High-quality genomic data could be used for Long-read sequencing and high-throughput chromosome (Hi-C) technology."
- The authors reference the benefits of Hi-C in their introduction but do not use Hi-C in their analysis.

Are sufficient details of methods and analysis provided to allow replication by others? (partly)

The authors provide the software and version numbers to software used which is appreciated and facilitates reproducibility. However, the authors do not specify if software was used with default parameters or parameters deviating from default. These can be critical for reproducing or evaluating the analyses done, especially for tools like BLAST. Nor do the authors provide links to repositories (e.g. github, zenodo) of scripts used for verification and/or reproduction of methods. I would strongly recommend a remark in the methods stating that unless otherwise specified, default parameters are used, and explicit description when non-standard parameters are used in text or a supplement. If possible, provide access to analysis scripts in their entirety.

Are the conclusions drawn adequately supported by the results? (Partly)

1. The assembly BUSCO assessments have a very low duplicate BUSCO count but the gene model level BUSCO assessments are much higher (~20%). This is not unexpected as isoform

redundancy, pseudo-genes, etc do crop up and can impact BUSCO analysis. Have the authors run the annotation level BUSCO analysis using only longest isoforms to verify? It is not clear from the methods how exactly BUSCO was run on the annotation results.

2. The authors do not mention any secondary verification on the gene models / predicted duplicates to verify they are not artifacts/pseudogenes/etc. While automated pipelines like funannotate are powerful and convenient, secondary validation of gene models is often needed especially if proceeding to duplication analysis with tools like MCScanX/OrthoFinder/Café.
3. The authors do not describe the assembly or annotation of the mitochondrial genome of each species. While not strictly necessary per se, these data are readily isolated from the long reads and would provide a more complete picture of the genomes and further context to the synteny-based phylogenetic analysis.
4. The authors use RagTag for reference-based scaffolding of assembled contigs. In the absence of Hi-C/OmniC data, this is a suitable alternative, especially given the authors have access to a closely related congener *Betta splendens*. The success of this approach is determined by the completeness/continuity/scaffold size of the input assemblies and the reference. Given this, comparisons of the assembly and annotation statistics of the new assemblies to the reference would provide important context for these assemblies. This would improve the ability of readers to gauge the quality of the assemblies and the success of the method. For example, a table produced by a tool like QUAST that compares the major assembly statistics would be suitable.
5. The RagTag method used can cause scaffolded assemblies to resemble the reference genome in terms of structure, especially for more fragmented/low coverage assemblies. The authors need to clearly discuss the potential impact of these potential artefacts on their collinearity results and conclusions.
6. Given the authors are using gene models for their comparative analysis (not sequences directly verified with molecular methods), the authors should report support for identified protein insertions with primary data (e.g. insertion spanning long reads, depth of coverage at the insertion site, etc.).
7. In the discussion, the authors make claims about purifying selection, rate of gene evolution, lineage specific selection, etc.. These are inferences that can only be drawn with population genetic data. These claims are improper for the data presented (e.g. assemblies of single individuals) and should be removed or qualified as such.

Additional comments

1. The manuscript would be strengthened by reporting metrics from initial k-mer profiling with GenomeScope2. These could be reported in the form of the annotated k-mer distribution plot generated by GenomeScope2.
2. The authors do not describe any methods for decontamination (either technical or biological). If these were not performed, this should be explicitly noted. If they were, what methods were employed and if they were pre-assembly or post assembly should be specified.
3. The authors do make publicly available the raw data, genome assemblies, and annotations. I would also recommend additionally submitting the repeats to public databases such as Dfam or FishTEDB.

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Partly

If applicable, is the statistical analysis and its interpretation appropriate?

I cannot comment. A qualified statistician is required.

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: bioinformatics, marine biology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 18 May 2026

Luthfi Nurhidayat

Comments

The study employs commonly used open-source tools and standard methodologies for assembly and annotation. The genomic resources generated by this study are novel and valuable contributions and will surely be useful for future studies in these species and for teleost and vertebrate genomics more broadly.

The manuscript would benefit from proof reading for spelling, grammar, consistency, etc.:

- Consistency of italicization for species names throughout.

Response:

We have checked and fixed the consistency of italicization for species names within the revised manuscript.

- Introduction: "The number of endemic wild Betta distribution and diversity of Southeast Asia has been identified for 16 years in Thailand (2 species) and Indonesia (9 species)"

Response:

We have revised it to "Indonesia harbors nine wild Betta species classified as Critically Endangered (CN) and seven species categorized as Endangered (EN). In comparison, Thailand contains two Critically Endangered (CE) wild Betta species and one Endangered (EN) species"

- Introduction: "Three species of *Betta* spp. were endemic, i.e., *Betta burdigala*, *Betta chloropharynx*, and *Betta schalleri*."

Response:

We have revised it to "Three *Betta* species—*B. burdigala*, *B. chloropharynx*, and *B. schalleri*—are endemic to the Bangka Islands"

- Introduction: "The genetic studies of the endemic fish *Betta* spp. on Bangka Island have..."

Response:

We have change it to "Previous studies on the endemic *Betta* species from Bangka Island have included genetic analyses based on mitochondrial DNA and eDNA metabarcoding, as well as comprehensive morphological examinations, all of which have contributed to clarifying the genetic relationships and taxonomic status of these species."

- Introduction: "However, the whole genome for the evolution of the endemic fish *Betta* spp. on Bangka Island hasn't been revealed."

Response:

We have revised it to "However, the whole-genome evolutionary landscape of the endemic *Betta* species from Bangka Island has not yet been explored."

- Introduction: "High-quality genomic data could be used for Long-read sequencing and high-throughput chromosome (Hi-C) technology."

Response:

Thank you for noticing the mistakes. We have revised it to "High-quality genomic data can be generated using long-read sequencing technologies such as Oxford Nanopore sequencing, which enables the production of highly contiguous genome assemblies....."

- The authors reference the benefits of Hi-C in their introduction but do not use Hi-C in their analysis.

Response:

We have revised the Introduction section by removing the Hi-C-related description to better align the manuscript with the actual sequencing and assembly strategy used in this study, which was based on Oxford Nanopore long-read sequencing without Hi-C data. The revised text now reads:

"High-quality genomic data can be generated using long-read sequencing technologies such as Oxford Nanopore sequencing, which enables the production of highly contiguous genome assemblies. These technologies help resolve complex repetitive regions and haplotype heterozygosity, both of which are major sources of genome assembly errors⁷."

This revision ensures that the Introduction more accurately reflects the methodologies and analyses performed in the present study.

Are sufficient details of methods and analysis provided to allow replication by others? (partly)

The authors provide the software and version numbers to software used which is appreciated and facilitates reproducibility. However, the authors do not specify if software was used with default parameters or parameters deviating from default. These can be critical for reproducing or evaluating the analyses done, especially for tools like BLAST. Nor do the authors provide links to repositories (e.g. github, zenodo) of scripts used for verification and/or reproduction of methods. I would strongly recommend a remark in the methods stating that unless otherwise specified, default parameters are used, and explicit

description when non-standard parameters are used in text or a supplement. If possible, provide access to analysis scripts in their entirety.

Response:

Thank you for this valuable suggestion. We agree that explicit reporting of software parameters and analytical workflows is essential for ensuring reproducibility and transparency. In response, we have revised the Methods section to clarify that, unless otherwise specified, all software tools were executed using their default parameters. We also added explicit descriptions for analyses in which non-default settings or filtering criteria were applied.

Specifically, we added the following sentence to clarify the BLAST settings used in the duplication analysis:

“The analyses were performed using an E-value threshold of $1e-3$, with results limited to a maximum of five target hits per query sequence and output generated in tabular format (outfmt 6). Computations were executed using 12 parallel threads, while all other parameters were retained at their default settings.”

Regarding the analysis scripts, most workflows in this study were conducted using established bioinformatics software through standard command-line execution rather than extensive custom scripting. Therefore, we focused on providing detailed descriptions of the analytical workflows, software versions, and key command parameters directly within the manuscript and supplementary materials to facilitate reproducibility. We also ensured that all non-default settings and critical filtering parameters are now explicitly documented in the revised manuscript.

We appreciate the reviewer’s recommendation, which helped improve the methodological clarity, transparency, and reproducibility of the study.

Are the conclusions drawn adequately supported by the results? (Partly)

1. The assembly BUSCO assessments have a very low duplicate BUSCO count but the gene model level BUSCO assessments are much higher (~20%). This is not unexpected as isoform redundancy, pseudo-genes, etc do crop up and can impact BUSCO analysis. Have the authors run the annotation level BUSCO analysis using only longest isoforms to verify? It is not clear from the methods how exactly BUSCO was run on the annotation results.

Response:

Thank you for this important observation. We agree that annotation-level BUSCO assessments can show elevated duplicated BUSCO proportions due to the inclusion of multiple transcript isoforms, fragmented predictions, or potential pseudogene annotations. In our study, the BUSCO assessment of the predicted gene models was performed using the complete predicted protein set generated from the annotation pipeline, including all annotated isoforms, rather than restricting the analysis to only the longest isoform per gene. We acknowledge that this approach may contribute to the higher duplicated BUSCO proportion observed at the annotation level compared with the genome assembly-level BUSCO assessment.

To clarify this point, we have revised the Methods section to explicitly describe how BUSCO was conducted on the annotation results, including the use of the full predicted protein dataset. We also added clarification in the Results that the elevated duplicated BUSCO values at the annotation level likely reflect transcript isoform redundancy and annotation complexity rather than extensive true biological duplication.

Nevertheless, the overall high completeness of the predicted protein sets still supports the

quality and comprehensiveness of the genome annotations generated in this study.

2. The authors do not mention any secondary verification on the gene models / predicted duplicates to verify they are not artifacts/pseudogenes/etc. While automated pipelines like funannotate are powerful and convenient, secondary validation of gene models is often needed especially if proceeding to duplication analysis with tools like MCScanX/OrthoFinder/Café.

Response:

Thank you for this important comment. We agree that automated genome annotation pipelines may occasionally retain fragmented gene models, pseudogenes, or assembly-related artifacts, and that additional validation can further improve confidence in downstream duplication analyses.

In our study, structural gene prediction was performed using GALBA within the BRAKER framework. The GALBA pipeline incorporates protein evidence from closely related species to improve gene prediction accuracy and reduce false-positive annotations compared with purely ab initio approaches. Subsequently, Funannotate was used for functional annotation processing, including integration of functional annotation results into the GFF3 files and generation of submission-ready annotation files.

The resulting annotations showed high completeness based on BUSCO assessment of the predicted protein sets, supporting the overall quality of the gene models used in downstream analyses. Consequently, the duplication analyses performed using MCScanX were based on curated predicted protein-coding genes generated from this evidence-supported annotation workflow.

Nevertheless, we acknowledge that some predicted duplicates may still include fragmented annotations or non-functional copies that cannot be conclusively distinguished without additional transcriptomic, population-level, or experimental validation. To address this concern, we revised the manuscript in the “Structural and Functional Annotation and Gene Duplication Analysis” subsection of the Results section by adding the following statement: “The identified duplication patterns should therefore be interpreted as comparative genomic patterns inferred from predicted gene models and homologous gene relationships, rather than as definitive evidence of fully functional duplicated genes.”

3. The authors do not describe the assembly or annotation of the mitochondrial genome of each species. While not strictly necessary per se, these data are readily isolated from the long reads and would provide a more complete picture of the genomes and further context to the synteny-based phylogenetic analysis.

Response:

Thank you for this insightful suggestion. We agree that mitochondrial genome assemblies can provide valuable complementary information for comparative and phylogenetic analyses. In the present study, however, our primary focus was on the nuclear genome assemblies and comparative analyses related to genome structure, synteny, and pigmentation-associated genes.

We have in fact successfully assembled the mitochondrial genomes of the studied *Betta* species from the long-read sequencing data. However, these mitochondrial genome datasets are currently being analyzed in a separate study addressing distinct biological questions related to mitochondrial evolution and phylogeography. To avoid overlap between manuscripts and maintain a clear scope for the present study, we therefore chose

not to include the mitochondrial genome analyses here.

Nevertheless, we acknowledge the value of mitochondrial genomic information and have added a statement in the manuscript clarifying that mitochondrial genome assemblies were recovered from the sequencing data and will be reported separately in a dedicated study.

4. The authors use RagTag for reference-based scaffolding of assembled contigs. In the absence of Hi-C/OmniC data, this is a suitable alternative, especially given the authors have access to a closely related congener *Betta splendens*. This success of this approach is determined by the completeness/continuity/scaffold size of the input assemblies and the reference. Given this, comparisons of the assembly and annotation statistics of the new assemblies to the reference would provide important context for these assemblies. This would improve the ability of readers to gauge the quality of the assemblies and the success of the method. For example, a table produced by a tool like QAST that compares the major assembly statistics would be suitable.

Response:

Thank you for this constructive suggestion. We agree that comparison of the newly generated assemblies against the reference genome provides important context for evaluating assembly quality and the effectiveness of the reference-guided scaffolding strategy.

we performed additional assembly quality assessment using QAST and have added the results as **Supplementary file 5. QAST assembly quality assessment statistics for the three *Betta* genome assemblies** in figshare (DOI:

<https://doi.org/10.6084/m9.figshare.31189546.v3>). The QAST report includes major assembly statistics such as contig number, total assembly size, N50/NG50, largest contig size, GC content, duplication ratio, gap content, and misassembly statistics, together with alignment-based comparisons against the reference genome.

The QAST assessment showed that all three assemblies possess high contiguity, with contig N50 values ranging from ~19.3–20.1 Mb and largest contigs exceeding 33 Mb. In addition, the assemblies exhibited very low numbers of detected misassemblies and low gap content, supporting the overall structural quality of the assemblies.

We believe the inclusion of the QAST assessment substantially improves the contextual evaluation and transparency of the genome assemblies presented in this study.

5. The RagTag method used can cause scaffolded assemblies to resemble the reference genome in terms of structure, especially for more fragmented/low coverage assemblies. The authors need to clearly discuss the potential impact of these potential artefacts on their collinearity results and conclusions.

Response:

Thank you for this important comment. We acknowledge that the use of reference-guided scaffolding may influence large-scale structural organization and potentially increase apparent similarity to the reference genome. In our study, however, the genome assemblies were initially generated *de novo* from Oxford Nanopore long reads, while RagTag was used only during the scaffolding stage to improve chromosome-scale continuity.

To address this concern, we have revised the Results and Discussion sections to more explicitly acknowledge the potential limitations associated with reference-guided scaffolding. We also clarified that the collinearity analyses were performed using annotated

protein-coding genes and orthologous gene order on the scaffolded assemblies, rather than direct whole-genome nucleotide alignments. Therefore, the syntenic patterns presented in this study primarily reflect conservation of gene content and relative gene order, rather than exact chromosome-scale structural identity.

In addition, we have moderated the wording throughout the manuscript to avoid overinterpretation. Statements implying definitive structural conservation or phylogenetic inference have been revised to more cautious descriptions such as “broadly conserved syntenic patterns” and “patterns generally consistent with established taxonomic relationships.” We now also explicitly note that these observations should be interpreted within the context of reference-guided scaffolding.

Nevertheless, because the assemblies themselves were generated *de novo* and the comparative analyses were based on orthologous gene annotations rather than direct reference-guided genome alignment, we believe the analyses still provide informative comparative insights into genome organization among the examined *Betta* species.

6. Given the authors are using gene models for their comparative analysis (not sequences directly verified with molecular methods), the authors should report support for identified protein insertions with primary data (e.g. insertion spanning long reads, depth of coverage at the insertion site, etc.). **Response:**

Thank you for this important suggestion. To validate the inferred insertion identified from the gene model, we mapped Oxford Nanopore long reads back to the assembled genome. Coverage analysis revealed consistent read depth across the insertion locus (~70–88×), with no evidence of local coverage collapse or assembly breakpoints, thereby supporting the authenticity of the insertion. In addition, long-read alignments spanning the insertion region were visually inspected using Integrative Genomics Viewer and have been included in the Supplementary Materials deposited in figshare (DOI:

<https://doi.org/10.6084/m9.figshare.31189546.v2>). The following supplementary materials have been added:

Supplementary Figure S3. Validation of the CSF1R1 insertion using Oxford Nanopore long-read alignments. Coverage depth across the insertion locus remained consistent (~70–88×), and multiple long reads continuously spanned the insertion region, supporting the authenticity of the predicted insertion and excluding local assembly artifacts.

Supplementary File S4. Coverage depth of the CSF1R1_1 region based on Oxford Nanopore long-read alignments.

7. In the discussion, the authors make claims about purifying selection, rate of gene evolution, lineage specific selection, etc.. These are inferences that can only be drawn with population genetic data. These claims are improper for the data presented (e.g. assemblies of single individuals) and should be removed or qualified as such.

Response:

Thank you for this valuable comment. We agree that strong evolutionary inferences such as purifying selection, lineage-specific selection, or adaptive evolution cannot be conclusively established from genome assemblies derived from single individuals without population-level genomic data. To avoid overinterpretation, we have revised the Discussion section to use more cautious and descriptive language that better reflects the scope of the current dataset.

The following changes were made in the revised manuscript:

- "...reflecting strong purifying selection on this essential signaling pathway" was revised to
"...reflecting high sequence conservation and potential evolutionary constraints on this essential signaling pathway."
- "Similarly, the non-concordant evolution of MC1R, despite its variable role in teleost pigmentation, suggests relaxed constraint or functional innovation within the *Betta* lineage." was revised to
"Similarly, the non-concordant evolution of MC1R, despite its variable role in teleost pigmentation, is consistent with high sequence divergence that may warrant future investigation into potential regulatory or functional divergence within the *Betta* lineage."
- "...may hint at ongoing lineage-specific regulatory or structural adaptations..." was revised to
"...may hint at lineage-specific regulatory or structural variations."
- "...potentially linked to ecological adaptations or evolutionary divergence..." was revised to
"...potentially linked to evolutionary divergence among these lineages. However, population-level genomic data will be required to definitively connect these structural variations to specific ecological adaptations."

These revisions were made throughout the Discussion section to ensure that the interpretations remain appropriately conservative and fully supported by the available data.

Additional comments

1. The manuscript would be strengthened by reporting metrics from initial k-mer profiling with GenomeScope2. These could be reported in the form of the annotated k-mer distribution plot generated by GenomeScope2.

Response:

Thank you for the suggestion. We have added the 21-mer profiling analysis to Figure 3. Accordingly, we have revised the figure caption as follows:

"Figure 3. 21-mer-based genome size estimation (upper panels) and summary of genome assembly characteristics (lower panels) for *Betta burdigala*, *Betta chloropharynx*, and *Betta schalleri*."

2. The authors do not describe any methods for decontamination (either technical or biological). If these were not performed, this should be explicitly noted. If they were, what methods were employed and if they were pre-assembly or post assembly should be specified.

Response:

Thank you for noticing. We apologize for not mentioning the contamination screening method used for our assembly. Prior to assembly, we conducted adapter contamination screening using Porechop_ABI, and post-assembly, we screened for foreign contamination using FCS-GX. We have now included the contamination screening details in the methods section under the "Genome Assembly and Annotation" subsection.

"Adapter contamination screening was performed using Porechop_ABI prior to assembly"
And "Foreign contamination was screened and removed using NCBI FCS-GX prior to scaffolding"

3. The authors do make publicly available the raw data, genome assemblies, and annotations. I would also recommend additionally submitting the repeats to public databases such as Dfam or FishTEDB.

Response:

Many thanks for the suggestion. We plan to submit it to the DFAM database for the near future.

Competing Interests: We declare that we do not have any competing interest.

Reviewer Report 26 February 2026

<https://doi.org/10.5256/f1000research.192703.r461103>

© 2026 Xing B. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Bingpeng Xing

Third Institute of Oceanography Ministry of Natural Resources, Xiamen, China

This manuscript focuses on three endemic Betta fishes from Bangka Island—*Betta burdigala*, *B. chloropharynx*, and *B. schalleri*—conducting an integrated study of meristic analysis and whole-genome assembly, and it makes the related sequencing data and assembly results publicly available. For endemic fishes of Bangka Island peatland wetlands, this study provides the first genomic resources and has potential value for conservation genomics and species evolution research.

Overall, the data release is relatively complete, and the assembly contiguity and BUSCO completeness are high, meeting the basic requirements of a resource-type genome paper. However, the manuscript still needs strengthening in the following aspects:

1. Reference genome-guided scaffolding may affect structural and collinearity conclusions. The authors used RagTag together with the *Betta splendens* reference genome for scaffold construction. This approach is feasible for generating chromosome-scale assemblies, but it may also introduce structural bias to some extent, making the assembly closer to the reference genome structure. The conclusions in the current manuscript such as “high collinearity,” “closer phylogenetic relationships,” and “highly conserved structure” are all based on this scaffolding result; therefore: it is recommended to clearly state the potential limitations of reference-guided scaffolding in the Methods and Discussion; and clarify whether the collinearity analysis was based on the original contigs or the reference-guided scaffolds.
2. The sample size is limited, and over-inference at the species level should be avoided. The morphological analysis includes only three individuals per species, and genome sequencing includes only one individual per species. This is acceptable for a resource-type paper, but inferences in the following aspects should be more cautious: patterns of interspecific differences; gene duplication dynamics; specific gene insertion events; and inference of evolutionary divergence time.
3. The two “large insertions” in CSF1RA require stronger supporting evidence. The manuscript

reports that CSF1RA in *B. burdigala* contains two large amino-acid insertions and speculates that they may affect pigment regulation. This result has potential novelty but currently lacks the following validation: long-read coverage; read-level support at the insertion boundaries; whether there are low-complexity or repetitive-sequence artifacts; and whether assembly or annotation errors have been excluded.

4. Evidence from a single-gene phylogenetic analysis is limited. The authors constructed a phylogenetic tree using the single gene CSF1RA and compared it with the NCBI taxonomy tree. A single-gene tree can only serve as a consistency check and cannot provide strong support for phylogenetic relationships.

5. Genome quality assessment metrics can still be strengthened. The manuscript reports N50, scaffold number, BUSCO, GC content, and gap percentage; these metrics are relatively basic. It is recommended to add sequencing depth (estimated coverage); read length N50 or mean read length; k-mer completeness analysis; assembly quality value (QV); and contamination detection methods (if any). Adding 2–3 of the above metrics can significantly improve confidence in the assembly.

6. The background statements in the Introduction needs revision. There are unclear and illogical sentences in the Introduction, such as the description of “16 years” and “2 species in Thailand and 9 species in Indonesia,” which is confusing. Rewrite the relevant paragraphs to make the logic clearer; describe more accurately recent progress in taxonomy and diversity research of the genus *Betta*; and standardize IUCN status as “Critically Endangered (CR)” or “Endangered (EN)” with clear citations.

7. The Latin name format is not standardized. The manuscript should be unified: genus and species names in italics, such as *Betta burdigala*; abbreviated form *B. burdigala*; in “*Betta* spp.” only *Betta* in italics and “spp.” not in italics; and the family name Osphronemidae not in italics. There are many cases of missing italics or inconsistency, and a full-text unified check is recommended.

8. There are many spelling and language errors, including but not limited to: critically → critically, Genome assembljies → assemblies, inconsistent spelling such as chlorpharynx / chloropharynx, etc. A comprehensive language check is recommended.

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?

Partly

If applicable, is the statistical analysis and its interpretation appropriate?

Partly

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: marine biology, genome, fish, genetic, DNA, eDNA

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 18 May 2026

Luthfi Nurhidayat

1. Reference genome-guided scaffolding may affect structural and collinearity conclusions. The authors used RagTag together with the *Betta splendens* reference genome for scaffold construction. This approach is feasible for generating chromosome-scale assemblies, but it may also introduce structural bias to some extent, making the assembly closer to the reference genome structure. The conclusions in the current manuscript such as “high collinearity,” “closer phylogenetic relationships,” and “highly conserved structure” are all based on this scaffolding result; therefore: it is recommended to clearly state the potential limitations of reference-guided scaffolding in the Methods and Discussion; and clarify whether the collinearity analysis was based on the original contigs or the reference-guided scaffolds.

Response:

Thank you for this thoughtful comment. We agree that reference-guided scaffolding may introduce structural bias toward the reference genome and that this limitation should be clearly acknowledged when interpreting large-scale synteny and collinearity patterns. In our study, the primary genome assemblies were first generated *de novo* from Oxford Nanopore long reads. The resulting contigs were subsequently scaffolded using RagTag with the *Betta splendens* genome as a guide to improve chromosome-scale continuity. We acknowledge that this scaffolding strategy may partially influence large-scale structural organization and therefore could increase apparent similarity to the reference genome. To clarify this point, we have revised the Result and Discussion sections to explicitly state the potential limitations associated with reference-guided scaffolding. We also clarified that the collinearity analyses were conducted using annotated protein-coding genes and orthologous gene order on the scaffolded assemblies, rather than direct whole-genome nucleotide alignments. Therefore, the inferred syntenic relationships primarily reflect conservation of gene content and relative gene order rather than exact chromosome-scale structural identity.

In addition, we have softened the interpretation throughout the Discussion section to avoid overstatement. For example, statements such as “significant genome conservation,” “highly conserved structure,” and “closer evolutionary relationship” were revised to more cautious wording including “broadly conserved syntenic patterns,” “comparatively higher levels of shared gene order,” and “patterns generally consistent with existing taxonomic relationships.” We also now emphasize that the observed collinearity patterns should be interpreted as comparative genomic signals within the context of reference-guided scaffolding, rather than definitive evidence of fully conserved chromosomal architecture. Nevertheless, because the assemblies were initially generated *de novo* and the synteny analysis was based on annotated orthologous genes rather than direct reference-guided

genome alignment, we believe the analyses still provide informative comparative insights into genome organization among the examined *Betta* species.

2. The sample size is limited, and over-inference at the species level should be avoided. The morphological analysis includes only three individuals per species, and genome sequencing includes only one individual per species. This is acceptable for a resource-type paper, but inferences in the following aspects should be more cautious: patterns of interspecific differences; gene duplication dynamics; specific gene insertion events; and inference of evolutionary divergence time.

Response:

Thank you for this important comment. We agree that the limited sample size, particularly the use of a single sequenced individual per species, requires cautious interpretation of species-level inferences. Our primary objective in this study was to generate foundational genomic resources for these threatened *Betta* species rather than to make definitive population-level evolutionary conclusions.

In response, we revised the manuscript to avoid overinterpretation by moderating statements related to interspecific differences, gene duplication patterns, insertion events, and evolutionary divergence. We replaced overly definitive wording with more cautious expressions such as “is consistent with,” “may represent,” or “suggests potential variation,” where appropriate.

We also added statements in the last paragraph of Discussion acknowledging that the assemblies represent single individuals and therefore may not capture the full extent of intraspecific genomic variation. Similarly, we noted that the limited number of specimens used for morphological analyses constrains broader species-level interpretations. However, we note that the examined *Betta* species are highly endemic and threatened taxa, making specimen collection ethically and logistically challenging. Within this context, the use of three specimens per species for morphological analyses and one individual per species for reference genome generation is consistent with many genomic resource studies involving rare or endangered organisms.

We believe these revisions provide a more balanced interpretation while preserving the value of the genomic resources presented in this study.

3. The two “large insertions” in CSF1RA require stronger supporting evidence. The manuscript reports that CSF1RA in *B. burdigala* contains two large amino-acid insertions and speculates that they may affect pigment regulation. This result has potential novelty but currently lacks the following validation: long-read coverage; read-level support at the insertion boundaries; whether there are low-complexity or repetitive-sequence artifacts; and whether assembly or annotation errors have been excluded.

Response:

Thank you for this important suggestion. To validate the inferred insertion identified from the gene model, we mapped Oxford Nanopore long reads back to the assembled genome. Coverage analysis revealed consistent read depth across the insertion locus (~70–88×), with no evidence of local coverage collapse or assembly breakpoints, thereby supporting the authenticity of the insertion. In addition, long-read alignments spanning the insertion region were visually inspected using Integrative Genomics Viewer and have been included in the Supplementary Materials deposited in figshare (DOI: <https://doi.org/10.6084/m9.figshare.31189546.v3>). The following supplementary materials

have been added:

Supplementary Figure S3. Validation of the CSF1R1 insertion using Oxford Nanopore long-read alignments. Coverage depth across the insertion locus remained consistent (~70–88×), and multiple long reads continuously spanned the insertion region, supporting the authenticity of the predicted insertion and excluding local assembly artifacts.

Supplementary File S4. Coverage depth of the CSF1R1_1 region based on Oxford Nanopore long-read alignments.

4. Evidence from a single-gene phylogenetic analysis is limited. The authors constructed a phylogenetic tree using the single gene CSF1RA and compared it with the NCBI taxonomy tree. A single-gene tree can only serve as a consistency check and cannot provide strong support for phylogenetic relationships.

Response:

Thank you for this valuable comment. We agree that phylogenetic inference based on a single gene has limited power for resolving species-level evolutionary relationships and should not be interpreted as definitive phylogenetic evidence. In our study, the CSF1RA phylogeny was intended primarily as a gene-level comparative analysis and a consistency check against the established taxonomy, rather than as a comprehensive reconstruction of species relationships. To avoid overinterpretation, we have revised the relevant sections of the manuscript to clarify the purpose and limitations of the single-gene analysis. We now explicitly state that the CSF1RA tree is presented to illustrate evolutionary conservation and relative clustering patterns among related taxa, whereas accepted species relationships remain based on established taxonomy and previous phylogenomic studies.

In the Results section, we revised the sentence:

“We found that the CSF1RA phylogenetic tree closely resembles the phylogenetic tree of the nine fish species compared based on NCBI taxonomic data”

to:

“Gene-based phylogenetic analyses showed that the CSF1RA tree exhibited clustering patterns broadly consistent with the established taxonomic relationships among the nine fish species based on NCBI taxonomy.”

We also added the following statement to the Results section:

“We note that these analyses are based on single genes and therefore are intended primarily for comparative assessment of gene-level evolutionary patterns rather than for robust species-level phylogenetic inference.”

In addition, we revised the Discussion section to reduce overinterpretation of the single-gene phylogenetic analyses and to present the evolutionary implications more cautiously.

5. Genome quality assessment metrics can still be strengthened. The manuscript reports N50, scaffold number, BUSCO, GC content, and gap percentage; these metrics are relatively basic. It is recommended to add sequencing depth (estimated coverage); read length N50 or mean read length; k-mer completeness analysis; assembly quality value (QV); and contamination detection methods (if any). Adding 2–3 of the above metrics can significantly improve confidence in the assembly.

Response:

Thank you for this valuable suggestion. We agree that including additional assembly quality metrics would strengthen confidence in the genome assemblies. In response, we have revised the manuscript to incorporate additional sequencing and quality-control

information.

First, we updated the *Genomic DNA extraction and sequencing* subsection in the Methods section. The original text:

"The genome sequence data for *Betta burdigala* includes 16.38 million reads and 37.39 gigabases. For *Betta chloropharynx*, the data comprises 14.97 million reads and 57.97 gigabases. Meanwhile, *Betta schalleri* has 11.83 million reads and 23.75 gigabases." has been revised to:

"The genome sequence data for *Betta burdigala* comprised 16.38 million reads totaling 37.39 Gb, with a read N50 of 5.23 kb, corresponding to an estimated genome coverage of approximately 74.8–83.1× assuming a genome size of 450–500 Mb. For *Betta chloropharynx*, sequencing generated 14.97 million reads totaling 57.97 Gb, with a read N50 of 7.75 kb, representing an estimated coverage of approximately 115.9–128.8×. Meanwhile, *Betta schalleri* produced 11.83 million reads totaling 23.75 Gb, with a read N50 of 5.5 kb, corresponding to an estimated genome coverage of approximately 47.5–52.8×."

These additions provide both sequencing depth estimates and read-length statistics requested by the reviewer.

Second, we updated the *Genome assembly and annotation* subsection to include contamination screening and preprocessing procedures that were inadvertently omitted from the original manuscript. Specifically, we added the following statements:

"Adapter contamination screening was performed using Porechop_ABI prior to assembly."

"Foreign contamination was screened and removed using NCBI FCS-GX prior to scaffolding."

Finally, we also incorporated the 21-mer-based genome size estimation analysis into the revised Figure 3, which provides an additional k-mer-based assessment supporting assembly completeness and consistency with the final assembly sizes.

6. The background statements in the Introduction needs revision. There are unclear and illogical sentences in the Introduction, such as the description of "16 years" and "2 species in Thailand and 9 species in Indonesia," which is confusing. Rewrite the relevant paragraphs to make the logic clearer; describe more accurately recent progress in taxonomy and diversity research of the genus *Betta*; and standardize IUCN status as "Critically Endangered (CR)" or "Endangered (EN)" with clear citations.

Response:

Thank you for the suggestion. We have revised the statement:

"The number of endemic wild *Betta* distribution and diversity of Southeast Asia has been identified for 16 years in Thailand (2 species) and Indonesia (9 species)¹⁻⁵" to

"Indonesia harbors nine wild *Betta* species classified as critically endangered and seven species categorized as endangered¹⁻³. In comparison, Thailand contains two critically endangered wild *Betta* species and one endangered species^{4,5}."

7. The Latin name format is not standardized. The manuscript should be unified: genus and species names in italics, such as *Betta burdigala*; abbreviated form *B. burdigala*; in "*Betta* spp." only *Betta* in italics and "spp." not in italics; and the family name Osphronemidae not in italics. There are many cases of missing italics or inconsistency, and a full-text unified check is recommended.

Response:

Thank you for this valuable comment, and we apologize for the inconsistencies in the

formatting of scientific names throughout the original manuscript. We agree that standardized formatting of Latin names is important for taxonomic clarity and overall manuscript readability.

In response, we conducted a full-text review of the manuscript to standardize the formatting of scientific names throughout the text, figures, tables, and supplementary materials. Genus and species names are now consistently italicized (e.g., *Betta burdigala*), abbreviated species names are formatted correctly (e.g., *B. burdigala*), and in expressions such as *Betta* spp., only the genus name is italicized while "spp." remains in standard font. In addition, family names such as Osphronemidae are now consistently presented in non-italicized format.

We also corrected multiple instances of missing italics and formatting inconsistencies throughout the manuscript to ensure conformity with standard zoological nomenclature conventions.

8. There are many spelling and language errors, including but not limited to: critically → critically, Genome assembljies → assemblies, inconsistent spelling such as chlorpharynx / chloropharynx, etc. A comprehensive language check is recommended.

Response:

Thank you for pointing this out. We appreciate the reviewer's careful reading of the manuscript. In response, we conducted a comprehensive language and spelling revision throughout the manuscript to correct typographical errors, improve grammar, and ensure consistency in terminology and species names.

Specifically, spelling errors such as "criticaly" to "critically," "Genome assembljies" to "genome assemblies," and inconsistent species naming (e.g., "chlorpharynx" to "chloropharynx") were corrected throughout the text. We also carefully reviewed formatting consistency, including the standardized use of italicization for scientific names.

Competing Interests: We declare that we do not have any competing interest.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research