

# The chromosome-level genome of an invasive Australian population of the eastern mosquitofish (*Gambusia holbrooki*)

Roger Huerlimann

`roger.huerlimann@oist.jp`

Okinawa Institute of Science and Technology Graduate University

Jennifer Donelson

James Cook University

Lucinda Aulsebrook

James Cook University

Timothy Ravasi

Okinawa Institute of Science and Technology Graduate University

---

data-descriptor

Keywords:

Posted Date: July 23rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10225135/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

---

# Abstract

The Eastern mosquitofish, *Gambusia holbrooki*, is a small livebearing fish native to the southeastern United States that has been introduced globally and is now considered one of the world's most invasive freshwater species. In Australia, established populations have undergone decades of adaptation to diverse thermal and ecological environments. Here we present a fully phased chromosome-level genome assembly of *G. holbrooki* from Townsville, Queensland, Australia, generated using PacBio HiFi long-read and Hi-C chromatin conformation sequencing. The assembly comprises 24 chromosomes and two haplotypes of approximately 676 Mb each, with contig and scaffold N50 values exceeding 26 Mb and 29 Mb, respectively. BUSCO completeness exceeded 98.4% for both haplotypes and Merquy quality values exceeded Q66. Telomeric repeat motifs were identified at both ends of 13 (HAP1) and 10 (HAP2) chromosome-scale scaffolds, with a further 9 (HAP1) and 7 (HAP2) chromosomes carrying signal at one end, indicating near-complete end-to-end assembly for most chromosomes. This resource provides a high-quality genomic reference for studying the invasion biology, local adaptation, sexual selection and evolutionary genomics of invasive fishes, and a foundation for future comparative analyses across globally distributed populations.

## Background & Summary

Eastern mosquitofish, *Gambusia holbrooki*, is a small (16–70 mm), tropical to subtropical freshwater poeciliid fish endemic to eastern North America (Pyke 2008). *G. holbrooki* and its sister species, the western mosquitofish, *Gambusia affinis*, were introduced globally as an ineffective biological control for malaria carrying mosquitos (Blaustein 1992), and are now highly invasive pest species on every continent other than Antarctica (Pyke 2008, Walton, Henke et al. 2012). *Gambusia spp.* were introduced to Europe, Asia, French Polynesia, the Caribbean, Africa and Australia in the 1920s (Pyke 2008, Walton, Henke et al. 2012). While there has been historical taxonomic uncertainty throughout this global colonisation, only *G. holbrooki* is located in Australia (Lloyd and Tomasov 1985).

Eastern mosquitofish are successful aquatic invaders globally due to their life history with rapid growth, early reproduction and short lifespan. Eastern mosquitofish mature rapidly, as fast as 18 to 20 days in both males and females (Meffe 1992). They have high reproductive investment, producing live young with high survival (Pyke 2008), and breed prolifically (generally 5–60 offspring per bout every 4–5 weeks; (Brown-Peterson and Peterson 1990, Meffe 1992, Jawad and Busneina 2000). Fertilisation is internal, with males using a modified anal fin (gonopodium) to transfer sperm to the female's gonoduct. Fecundity is related to female body size and juveniles are well developed at birth (6–8 mm; Meffe 1990). Female *Gambusia spp.* can also store viable sperm for long periods and may adjust when their eggs are fertilized depending on environmental conditions (Krumholz 1948, Haynes 1993). *Gambusia spp.* form size influenced hierarchies in which females are dominant and rank position is enforced by aggression (Bisazza and Marin 1991, Mills, Rader et al. 2004, Chen, Beekman et al. 2010, Matthews and Wong 2015).

The invasion success of eastern mosquitofish is also due to their broad ecological and physiological niches (e.g. temperature, hypoxia, salinity, pH (Carter and Wilson 2006, Stoffels, Weatherman et al. 2017, Kurtul, Tarkan et al. 2024). They have both a broad optimal thermal performance range (26–32°C; Wilson 2005, Pyke 2008) and reproduction thermal range (18–34°C; Wilson and Johnston 2004). While temperatures of 32°C and above are considered stressful (Meffe 1992, Matthews and Wong 2015, Aulsebrook, Hosler et al. 2026), they can survive in aquatic systems with temperatures between 10 and 35°C (Meffe 1991, Walton, Henke et al. 2012). Eastern mosquitofish are also highly flexible in diet (Rees 1958 as reviewed by Pyke 2008, Specziár 2004), and plastic in key life history traits (e.g., size at maturity, growth rate, fecundity) and physiological performance traits (e.g., swimming speeds, metabolism) to environmental variation (e.g., temperature, salinity; Meffe 1992, Alcaraz and García-Berthou 2007, Loughland and Seebacher 2020). Rapid physiological plasticity is possible within weeks of environmental exposure (Meffe 1992, Loughland and Seebacher 2020).

The eastern mosquitofish is also a model species for studying sexual selection, due to male-male competition (Booksmythe, Backwell et al. 2013, Harrison, Vega-Trejo et al. 2023), female mate choice (Bisazza, Vaccari et al. 2001, Kahn, Mautz et al. 2009), male mate choice (Agrillo, Dadda et al. 2008, Wong and McCarthy 2009), coercive mating (Wilson 2005), and sperm competition (Zane, Jones et al. 1999). Winners of male competition tend to be larger and have better pre-copulatory outcomes (Booksmythe, Backwell et al. 2013, Harrison, Vega-Trejo et al. 2023) including blocking smaller males from access to females (Bisazza and Marin 1995). There is evidence that females preferentially associate and mate with certain males based on growth rate and body size (Kahn, Livingston et al. 2012, Kim, Moran et al. 2021), gonopodium length (Kahn, Mautz et al. 2009), male behaviour (Bisazza and Marin 1995) and past mating experience (Aich, Jennions et al. 2020); however, coercive mating is also prominent (Wilson 2005, Condon and Wilson 2006). Males incessantly pursue females and attempt to mate (Wilson 2005) and this sexual conflict can lower female foraging efficiency and reproductive success (Pilastro, Benetton et al. 2003, Iglesias-Carrasco, Fox et al. 2019). Females mate multiply, and sperm competition is intense (Zane, Jones et al. 1999). The high costs of sperm production are indicated by the fact it is condition-dependent (O'dea, Jennions et al. 2014) and modified depending on the social environment (Wong and McCarthy 2009).

In this study, we report a fully phased chromosome-level genome assembly for *Gambusia holbrooki* generated using PacBio HiFi long-read and Hi-C chromatin conformation sequencing. The assembly comprised two haplotypes of approximately 676 Mb each, with contig and scaffold N50 values exceeding 26 Mb and 29 Mb, respectively. A total of 24 chromosome-scale scaffolds spanning 667.8 Mb (98.8%) of the haplotype 1 assembly were recovered, with only 8.3 Mb remaining as unplaced scaffolds. Telomeric repeat motifs were identified at both ends of 13 (HAP1) and 11 (HAP2) chromosome-scale scaffolds, with a further 7 and 8 chromosomes, respectively, carrying signal at only one end, indicating near-complete end-to-end assembly for the majority of chromosomes. BUSCO completeness exceeded 98.4% for both haplotypes, while Merquy quality values exceeded Q67. Structural annotation using EGAPX identified high-quality gene models with annotation BUSCO completeness of 99.3%. This

assembly provides a valuable genomic resource for studies of invasion biology, local adaptation, and evolutionary genomics in invasive freshwater fishes.

## Methods

### Ethics statement

Experimental procedures were approved by the James Cook University Animal Ethics Committee (JCU A2930) and were performed in accordance with the Australian Code for the Care and Use of Animals for Scientific Purposes. Wild fish were collected and used under general fisheries permit 270220 and biosecurity permit PRID001014 from the Queensland Government Department of Agriculture and Fisheries.

### Sample collection and nucleic acid extraction

A male *Gambusia holbrooki* adult was collected from Waterway Drive, Townsville, Australia (-19.423, 146.698) during June 2024 and brought back to the Marine and Aquaculture Research Facility at James Cook University, Townsville Australia where it was treated for potential disease or parasites with formalin (250 ppt) for 45 minutes before being placed in a 10 L tank within the laboratory. The male was kept at 27°C. In March 2025, the chosen fish was euthanized in an ice bath and immersed in liquid nitrogen immediately after removing it from the ice bath. At this time the fish was approximately 30 mm in length and 0.2 g in weight. The sample was stored at -80°C before shipping to the Okinawa Institute of Science and Technology on dry ice. The different extractions and library preparations were carried out at the OIST Sequencing Center.

The fish was kept frozen and cut into two pieces just past the anal pore (Fig. 1). The anterior part was used for RNA extraction using a Direct-zol RNA Kit (R2061, Zymo Research). The posterior part was subsampled for HMW DNA extraction using innuPREP Plant DNA I Kit (845-IPP-1516016, IST Innuscreen GmbH), and Chromatin Crosslinking using the Disuccinimidyl Glutarate and Formaldehyde. RNA quality control was carried out on a TapeStation 4200 using a RNA ScreenTape (5067–5576, Agilent). HMW DNA quality control was carried out on a Femto Pulse Genomic DNA 165 kb kit (FP-1002, Agilent).

### Sequencing

RNAseq library preparation was carried out using a NEBNext Ultra II Directional RNA Library Prep Kit. HiC library preparation was carried out using a Dovetail® Omni-C Kit. The libraries were pooled and sequenced on a single lane of an Illumina NovaSeq X Plus platform (San Diego, CA, USA) 1.5B 300-cycle flow cell.

PacBio HiFi library preparation was carried out with a SMRTbell® Prep Kit 3.0 (102-141-700, PacBio) and sequenced on one Revio SMRT Cell (102-202-200, PacBio) on a PacBio Revio system (102-090-600,

PacBio).

## Genome assembly and scaffolding

Genome size and heterozygosity of *Gambusia holbrooki* were estimated from PacBio HiFi reads using KAT (v2.4.2; Mapleson, Garcia Accinelli et al. 2017) through 21-mer frequency analysis. K-mer spectra analysis identified a dominant peak at approximately 122× coverage and a secondary heterozygous peak at approximately 63× coverage, consistent with a moderately heterozygous diploid genome. The genome size was estimated at 589.97 Mb, with an estimated heterozygosity rate of 0.70%.

Genome assembly was performed using hifiasm (v0.25.0; Cheng, Concepcion et al. 2021) with default parameters, incorporating PacBio HiFi reads and Hi-C data. The haplotype-resolved assembly resulted in two phased haplotypes designated HAP1 and HAP2 comprising 105 contigs with total assembly sizes of 676.1 Mb and 674.3 Mb, and contig N50 values of 26.0 Mb and 26.6 Mb, respectively.

Hi-C reads were mapped to each haplotype assembly independently using bwa-mem2 (v2.2.1, Vasimuddin, Misra et al. 2019) with the `-5SP` flags to retain split and supplementary alignments appropriate for Hi-C data. Resulting SAM files were converted to BAM format using SAMtools (v1.23.1, Danecek, Bonfield et al. 2021), with secondary and unmapped reads filtered using the `-F 2316` flag, and sorted by read name prior to downstream processing. Hi-C contact pairs were parsed, sorted, and deduplicated using pairtools, applying a minimum mapping quality threshold of 40 (`--min-mapq 40`), a walks policy retaining only uniquely mapping reads (`--walks-policy 5unique`), and a maximum inter-alignment gap of 30 bp. Valid contact pairs were filtered to retain uniquely mapped (UU), uniquely-rescued (UR), and rescued-unique (RU) pair types. Chromosome-scale scaffolding was performed with YaHS (v1.2.2, Zhou, McCarthy et al. 2023)) using a minimum alignment quality of 10. Contact maps were generated using juicer pre from the YaHS output AGP file and converted to .hic format using Juicer Tools (v1.22.01, Durand, Shamim et al. 2016) for visual inspection of scaffolding quality. Both phases were then screened for adapter and vector contamination prior to submission using NCBI FCS-adaptor (Astashyn, Tvedte et al. 2024). Assembly quality was assessed at both the contig and scaffold stages using QUAST (v5.0.2, Gurevich, Saveliev et al. 2013) for assembly statistics, BUSCO (v5.1.3, Manni, Berkeley et al. 2021) against the actinopterygii\_odb10 lineage dataset (3,640 ortholog) for gene completeness, and Merqury (v1.4, Rhie, Walenz et al. 2020) using a k-mer size of 21 derived from the HiFi reads for reference-free base accuracy and completeness estimation. Hi-C scaffolding of the phased assemblies produced chromosome-scale HAP1 and HAP2 assemblies with total sizes of 676.1 Mb and 674.3 Mb, comprising 92 and 105 scaffolds and scaffold N50 values of 30.2 Mb and 29.2 Mb, respectively. The 24 largest scaffolds corresponded to the expected chromosome number and accounted for 98.8% and 98.1% of the total assembly length in HAP1 and HAP2, respectively. Merqury k-mer analysis indicated high base-level accuracy for both haplotypes, with quality values of Q68.0 and Q66.8 and k-mer completeness of 96.4% and 96.2% for HAP1 and HAP2, respectively.

Telomeric repeat content was assessed with *tidk* (v0.2.65, Brown, Manuel Gonzalez de La Rosa et al. 2025) using the canonical vertebrate telomeric repeat unit (TTAGGG) and a window size of 10,000 bp, with a threshold of  $\geq 50$  repeats per window used to distinguish strong from weak signal. Of the 24 chromosome-scale scaffolds, 13 (HAP1) and 10 (HAP2) carried a strong telomeric signal at both ends, a further 9 (HAP1) and 7 (HAP2) carried strong or weak signal at one end paired with weak signal at the other, and the remaining 2 (HAP1) and 7 (HAP2) chromosomes showed either no detectable signal or only weak signal at one or both ends, indicating near-complete end-to-end assembly for the majority of chromosomes in both haplotypes.

The mitochondrial genome was assembled and annotated from HiFi contigs using MitoHiFi (v3.2.1, Uliano-Silva, Ferreira et al. 2023), with the *Gambusia affinis* mitogenome (NC\_028274.1) as the reference for contig identification and annotation transfer. MITOS2 (v2.0.2, Bernt, Donath et al. 2013) was used for de novo annotation (--mitos), and genetic code 2 (vertebrate mitochondrial) was specified (-o 2).

## Genome annotation

Repeat annotation and masking of both assemblies was performed using a multi-round strategy combining de novo repeat discovery with library-based masking. A species-specific repeat library was first constructed using RepeatModeler2 with LTR structural detection enabled (-LTRStruct). The resulting repeat families were prefixed with a species tag and partitioned into known and unknown elements. Unknown repeats were classified in two successive rounds using reclassifier (<https://github.com/darencard/GenomeAnnotation/blob/master/reclassifier>): first against *Gambusia*-specific entries from the Dfam 3.9 database (Storer, Hubley et al. 2021) combined with the RepeatModeler-identified known elements as reference, and second in a self-referential round using the classified output from round one. The final curated library comprised classified known elements and residual unknown elements.

Repeat masking was then applied in four sequential RepeatMasker rounds, each operating on the softmasked output of the previous round: (1) simple and low-complexity repeats; (2) *Gambusia* transposable elements sourced from Dfam 3.9; (3) known elements from the species-specific de novo library; and (4) residual unknown elements from the de novo library. Outputs from all four rounds were merged and re-summarised using ProcessRepeats. Repeat annotations were converted to GFF3 format using rmOutToGFF3, and repeat composition per subfamily was tabulated as a proportion of total genome length.

Structural gene annotation of both haplotype assemblies was performed using the NCBI Eukaryotic Genome Annotation Pipeline (EGAPx v0.5.2, Nawrocki and Infernal, Griffiths-Jones, Bateman et al. 2003, Dobin, Davis et al. 2013, Shen, Le et al. 2016, Buchfink, Reuter et al. 2021, Chan, Lin et al. 2021, Danecek, Bonfield et al. 2021, Li 2023) executed via Nextflow (v25.10.4, Di Tommaso, Chatzou et al. 2017) and Singularity (v4.1.4, Kurtzer, Sochat et al. 2017). Support data included the RNA-seq reads produced in this study and EGAPx provided data associated with taxid 27720. Support data were pre-downloaded to

a local cache prior to running the annotation. Submission files were prepared from the EGAPx annotated genome ASN output using the `prepare_submission` tool as separate `.sqn` submissions files for HAP1 (locus tag prefix AC2TIX) and HAP2 (locus tag prefix AC2TIY).

The EGAPx annotation of the HAP1 and HAP2 assemblies predicted 24,433 and 24,381 genes, comprising 22,780 and 22,695 protein-coding genes, respectively, alongside non-coding genes, pseudogenes, and immunoglobulin/T-cell receptor segment genes. BUSCO completeness of the annotated gene sets reached 99.3% and 99.0% for HAP1 and HAP2, respectively, against the *actinopterygii\_odb10* lineage dataset. Additional functional annotation was performed using the eggNOG-mapper (v2.1.12, Cantalapiedra, Hernández-Plaza et al. 2021) singularity container against the eggNOG orthology database (v5.0, Huerta-Cepas, Szklarczyk et al. 2019), with Diamond (Buchfink, Xie et al. 2015) as the aligner.

## Genome comparison

Whole-genome alignments were performed using `nf-core/pairgenomealign` v3.5.2 (Plessy, Mansfield et al. 2024) run via Nextflow (v25.10.4), which implements pairwise genome alignment with LAST (Kiełbasa, Wan et al. 2011), reduction to a best one-to-one alignment set using `last-split` (Frith and Kawaguchi 2015), and dot-plot visualization with `last-dotplot`. HAP1 was used as the alignment target, with HAP2, the ERGA-BGE *G. holbrooki* (Italy) assembly (GCA\_965282415.1; Fraser et al. 2026), and the *G. affinis* assembly (GCF\_019740435.1; Shao et al. 2020) as queries. Alignments were run using the RY128 seeding scheme (`--seed RY128`) and filtered to retain only one-to-one alignment blocks (`--dotplot_filter`).

## Data Records

Raw PacBio HiFi, Hi-C, and RNA-seq sequencing data are available in the NCBI Sequence Read Archive under BioProject PRJNA1450840, Study accession SRP690386 (<https://identifiers.org/ncbi/insdc.sra:SRP690386>), with individual run accessions SRR38011946 (HiFi), SRR38065502 (Hi-C), and SRR38065501 (RNA-seq). The annotated haplotype 1 (HAP1) and haplotype 2 (HAP2) genome assemblies, generated using EGAPx, are currently available on Zenodo (10.5281/zenodo.20725785) and are under processing at Genbank. The Genbank accession numbers will be added as soon as they become available.

Additional genome assembly and annotation outputs (repeat content, EggNog outputs, various statistics) are available at 10.5281/zenodo.20725785.

## Technical Validation

## Assembly quality

Genome size and heterozygosity were estimated from the raw PacBio HiFi reads by 21-mer frequency analysis using KAT, yielding an estimated haploid genome size of 589.97 Mb and a heterozygosity rate of 0.70% (Fig. 2). The k-mer spectra showed a clean bimodal distribution with a heterozygous peak at ~ 68X and a homozygous peak at ~ 123X coverage, consistent with a moderately heterozygous diploid genome and high-quality HiFi data with no evidence of contamination or unexpected ploidy.

Assembly statistics for both haplotypes are summarised in Table 1. Both haplotype assemblies achieved high contiguity, with contig N50 values of 26.0 Mb (HAP1) and 26.6 Mb (HAP2) prior to scaffolding. Following Hi-C scaffolding with YaHS, scaffold N50 values increased to 30.2 Mb and 29.2 Mb for HAP1 and HAP2, respectively. The 24 largest scaffolds accounted for 667.8 Mb (98.8%) and 661.6 Mb (98.1%) of the total assembly length in HAP1 and HAP2, corresponding to the expected chromosome number of  $2n = 48$  for *Gambusia holbrooki*. Base-level accuracy assessed using Merquy yielded quality values of QV68.0 and QV66.8 for HAP1 and HAP2, respectively, with k-mer completeness of 96.4% and 96.2%, indicating exceptionally high sequence accuracy consistent with PacBio HiFi data. BUSCO completeness against the actinopterygii\_odb10 lineage dataset ( $n = 3,640$ ) was 98.6% and 98.4% for HAP1 and HAP2 scaffolds, with negligible change from the contig-level scores (98.6% and 98.4%), confirming that scaffolding did not introduce gene-space errors. Total gap content introduced by scaffolding was minimal at 3,700 bp (37 joins) for HAP1 and 3,300 bp (33 joins) for HAP2, with four and six chromosomes assembled as single unjoined contigs in HAP1 and HAP2, respectively.

Table 1  
Assembly and annotation statistics for *Gambusia holbrooki* HAP1 and HAP2

| <b>Metric</b>                  | <b>HAP1</b>      | <b>HAP2</b>      |
|--------------------------------|------------------|------------------|
| Total assembly length          | 676.0 Mb         | 674.2 Mb         |
| Chromosome-scale sequence      | 667.8 Mb (98.8%) | 661.6 Mb (98.1%) |
| Number of contigs/scaffolds    | 105 / 92         | 105 / 105        |
| Number of chromosomes          | 24               | 24               |
| Contig N50                     | 26.0 Mb          | 26.6 Mb          |
| Scaffold N50                   | 30.2 Mb          | 29.2 Mb          |
| GC content                     | 39.0%            | 39.0%            |
| N's per 100 kbp                | 0.55             | 0.49             |
| BUSCO completeness (scaffolds) | 98.6%            | 98.4%            |
| BUSCO completeness (proteins)  | 99.3%            | 99.0%            |
| Mercury QV                     | 68.0             | 66.8             |
| Mercury completeness           | 96.4%            | 96.2%            |
| Repeat content                 | 29.2%            | 28.7%            |
| Protein-coding genes           | 22,780           | 22,695           |
| Total genes                    | 24,433           | 24,381           |
| Proteins with EggNOG hit       | 98.5%            | 98.5%            |

Hi-C contact maps generated using HiCExplorer confirmed high-quality chromosomal scaffolding for both haplotypes, with strong intra-chromosomal signal along the diagonal and negligible inter-chromosomal contacts across all 24 chromosome-scale scaffolds (Fig. 3). Telomeric repeat analysis using tidk identified the canonical vertebrate repeat unit (TTAGGG) at chromosome termini. 13 of 24 HAP1 chromosomes and 10 of 24 HAP2 chromosomes showed strong bilateral telomeric signal ( $\geq 50$  repeats per 10 kb window), with an additional 9 (HAP1) and 7 (HAP2) chromosomes showing strong or weak signal at one end, 1 (HAP1) and 3 (HAP2) showing strong signal at only one end with no detectable signal at the other, and 1 (HAP1) and 4 (HAP2) showing weak or absent signal at both ends.

The contig and scaffold size distributions for both haplotypes further illustrate the high contiguity of the assembly and the effectiveness of Hi-C scaffolding (Fig. 4). Prior to scaffolding, both HAP1 and HAP2 comprised 105 contigs each, with a gradual size distribution extending down to small unplaced fragments. Following scaffolding, 24 chromosome-scale sequences are clearly distinguishable as a distinct group of large sequences to the left of the red dashed line, with a sharp drop-off to the remaining unplaced scaffolds, particularly pronounced in HAP1, which was reduced from 105 contigs to 92

scaffolds. The step change between chromosomal and unplaced sequences is more abrupt in the scaffold panels than in the contig panels for both haplotypes, reflecting successful ordering and orientation of contigs into chromosome-scale scaffolds by YaHS.

## Repeat and gene annotation quality

Repeat annotation identified a total repeat content of 29.17% of the HAP1 assembly, consistent with estimates for other teleost fish genomes. DNA transposons were the most abundant repeat class (12.91%), followed by unclassified de novo elements (5.85%) and LINEs (4.06%).

Structural annotation of the HAP1 and HAP2 assemblies predicted 24,433 and 24,381 genes respectively, of which 22,780 and 22,695 were protein-coding, within the expected range for teleost fish. BUSCO assessment of the predicted protein sets reached 99.3% (HAP1) and 99.0% (HAP2) completeness, exceeding the genome-level BUSCO scores and confirming that EGAPx successfully annotated nearly all conserved genes present in the assembly. Of the 43,928 HAP1 protein sequences submitted to EggNOG-mapper, 98.5% received an orthology hit, 73.9% were assigned GO terms, and 72.5% received KEGG KO assignments, indicating comprehensive functional coverage.

## Assembly comparison against congeneric and conspecific genomes

To assess the assembly for missassemblies, whole-genome alignments were performed between the HAP1 assembly and (i) the second haplotype of this assembly (HAP2), (ii) the existing chromosome-level *G. holbrooki* reference assembly (ERGA-BGE, GCA\_965282415.1; Fraser, Gasparini et al. 2026), and (iii) the chromosome-level *G. affinis* assembly (GCF\_019740435.1; Shao, Ludwig et al. 2020). Alignments were visualized as one-to-one (O2O) dot plots (Suppl. Figures 1–3).

Self-alignment of HAP1 against HAP2 showed complete one-to-one correspondence across 23 chromosomes (Suppl. Figure 1). One inversion was observed near the end of chromosome 24, the only departure from otherwise complete collinearity between haplotypes.

Alignment of HAP1 against the existing *G. holbrooki* reference assembly likewise showed strong one-to-one synteny across 22 chromosomes (Suppl. Figure 2). One inversion each was found on chromosomes 23 and 24, with the Chr24 inversion mirroring the one found between HAP1 and HAP2.

Alignment against the *G. affinis* assembly showed the same one-to-one chromosomal correspondence and conserved synteny, consistent with the close phylogenetic relationship between the two congeneric species (Suppl. Figure 3). The exceptions were Chr06, Chr23, and Chr24, which all showed an inversion. The inversions on Chr23 and Chr24 are similar to the ones found in the alignment against the *G. holbrooki* (Italy) assembly.

The consistent appearance of the chr23 and chr24 inversions within this study's own haplotype pair and the two other alignments, suggests these may represent genuine, possibly conserved structural features

rather than assembly artifacts; further characterization of these regions was beyond the scope of this study.

Beyond synteny, the three assemblies showed close agreement across core assembly and annotation metrics (Table 2). Assembly length was similar across all three genomes (HAP1: 676.0 Mb; *G. holbrooki* ERGA: 674.6 Mb; *G. affinis*: 679.4 Mb), as was scaffold N50 (30.2 Mb, 29.6 Mb, and 29.8 Mb, respectively). Protein-coding gene counts were likewise comparable (22,780, 21,050, and 23,997, respectively), falling within the expected range for teleost fish genomes.

This consistency across independently generated, chromosome-scale assemblies of two closely related species further support the structural completeness and accuracy of the present assembly.

Table 2

Comparison of key assembly and annotation metrics between the HAP1 assembly (this study), the published *Gambusia holbrooki* (Italy) and *Gambusia affinis* assemblies.

| Metric                      | <i>G. holbrooki</i><br>Australia<br>(this study) | <i>G. holbrooki</i><br>Italy<br>GCA_965282415.1<br>(Fraser 2026) | <i>G. affinis</i><br>China<br>GCF_019740435.1<br>(Shao 2020) |
|-----------------------------|--------------------------------------------------|------------------------------------------------------------------|--------------------------------------------------------------|
| Assembly length             | 676.0 Mb                                         | 674.6 Mb                                                         | 679.4 Mb                                                     |
| Chromosomes                 | 24                                               | 24                                                               | 24                                                           |
| Scaffold N50                | 30.2 Mb                                          | 29.6 Mb                                                          | 29.8 Mb                                                      |
| BUSCO completeness (genome) | 98.6%                                            | 99.3%                                                            | 97.2%                                                        |
| Protein-coding genes        | 22,780                                           | 21,050                                                           | 23,997                                                       |
| Sequencing / scaffolding    | PacBio HiFi + Hi-C                               | PacBio HiFi + Hi-C                                               | ONT + BioNano + Hi-C                                         |

## Declarations

### Data Records

All sequencing data, genome assemblies, and annotation outputs generated in this study, as described in the Data Records section, are available under NCBI BioProject PRJNA1450840 and on Zenodo (10.5281/zenodo.20725785).

### Code Availability

The presented data analysis mostly used standard bioinformatic tools using standard parameters as suggested by the bioinformatic tools. Custom scripts were used in some places, mostly for data

handling. The complete code can be found under [https://github.com/R-Huerlimann/Australian\\_mosquitofish\\_genome/](https://github.com/R-Huerlimann/Australian_mosquitofish_genome/) (10.5281/zenodo.20659009)

## Funding

This work was supported by the Australian Research Council (J.M.D. and T.R.: DP230100078) and the Okinawa Institute of Science and Technology Graduate University (OIST).

## Author Contribution

RH: Formal analysis, Data curation, Software (bioinformatics), Visualization, Writing – original draft, Writing – review & editing. LA: Methodology (fish husbandry and experimental execution), Project administration, Writing – review & editing. JD and TR: Conceptualization, Funding acquisition, Project administration, Supervision, Writing – review & editing.

## Acknowledgement

We are grateful for the help and support provided by the OIST Sequencing Section, especially Mayumi Kawamitsu who carried out the extractions, library preparation and sequencing, and the Scientific Computing Section of the Research Support Division at Okinawa Institute of Science and Technology Graduate University. We are also grateful to the James Cook University Marine and Aquaculture Research Facility for support in the maintenance of the aquarium system. Thanks to Taylor Hosler and Cheuk Kay (Kate) Chu for support with animal care, and to Jeffrey Jolly and Yuno Kaneshi of the Marine Climate Change unit for all their help with the shipping logistics.

## Data Availability

Raw PacBio HiFi, Hi-C, and RNA-seq sequencing data are available in the NCBI Sequence Read Archive under BioProject PRJNA1450840, Study accession SRP690386 (<https://identifiers.org/ncbi/insdc.sra:SRP690386>), with individual run accessions SRR38011946 (HiFi), SRR38065502 (Hi-C), and SRR38065501 (RNA-seq). The annotated haplotype 1 (HAP1) and haplotype 2 (HAP2) genome assemblies, generated using EGAPx, are currently available on Zenodo (10.5281/zenodo.20725785) and are under processing at Genbank. The Genbank accession numbers will be added as soon as they become available. Additional genome assembly and annotation outputs (repeat content, EggNog outputs, various statistics) are available at 10.5281/zenodo.20725785.

## References

1. Agrillo, C., et al. (2008). "Choice of female groups by male mosquitofish (*Gambusia holbrooki*)."  
Ethology **114**(5): 479-488.
2. Aich, U., et al. (2020). "An experimental test of the role of male mating history on paternal effects in the livebearer fish *Gambusia holbrooki*." Biology Letters **16**(2): 20190945.
3. Alcaraz, C. and E. García-Berthou (2007). "Life history variation of invasive mosquitofish (*Gambusia holbrooki*) along a salinity gradient." Biological Conservation **139**(1-2): 83-92.
4. Astashyn, A., et al. (2024). "Rapid and sensitive detection of genome contamination at scale with FCS-GX." Genome biology **25**(1): 60.
5. Aulsebrook, L. C., et al. (2026). "Developmental warming induces severe deformities and mortality in a thermally tolerant fish species." Scientific Reports.
6. Bernt, M., et al. (2013). "MITOS: improved de novo metazoan mitochondrial genome annotation." Molecular phylogenetics and evolution **69**(2): 313-319.
7. Bisazza, A. and G. Marin (1991). "Male size and female mate choice in the eastern mosquitofish (*Gambusia holbrooki*: Poeciliidae)." Copeia: 730-735.
8. Bisazza, A. and G. Marin (1995). "Sexual selection and sexual size dimorphism in the eastern mosquitofish *Gambusia holbrooki* (Pisces Poeciliidae)." Ethology Ecology & Evolution **7**(2): 169-183.
9. Bisazza, A., et al. (2001). "Female mate choice in a mating system dominated by male sexual coercion." Behavioral Ecology **12**(1): 59-64.
10. Blaustein, L. (1992). "Larvivorous fishes fail to control mosquitoes in experimental rice plots." Hydrobiologia **232**(3): 219-232.
11. Booksmythe, I., et al. (2013). "Competitor size, male mating success and mate choice in eastern mosquitofish, *Gambusia holbrooki*." Animal Behaviour **85**(2): 371-375.
12. Brown-Peterson, N. and M. S. Peterson (1990). "Comparative life history of female mosquito fish, *Gambusia affinis*, in tidal freshwater and oligohaline habitats." Environmental Biology of Fishes **27**(1): 33-41.
13. Brown, M. R., et al. (2025). "Tidk: a toolkit to rapidly identify telomeric repeats from genomic datasets." Bioinformatics **41**(2): btaf049.
14. Buchfink, B., et al. (2021). "Sensitive protein alignments at tree-of-life scale using DIAMOND." Nature methods **18**(4): 366-368.
15. Buchfink, B., et al. (2015). "Fast and sensitive protein alignment using DIAMOND." Nature methods **12**(1): 59-60.
16. Cantalapiedra, C. P., et al. (2021). "eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale." Molecular biology and evolution **38**(12): 5825-5829.
17. Carter, A. J. and R. S. Wilson (2006). "Improving sneaky-sex in a low oxygen environment: reproductive and physiological responses of male mosquito fish to chronic hypoxia." Journal of experimental biology **209**(24): 4878-4884.

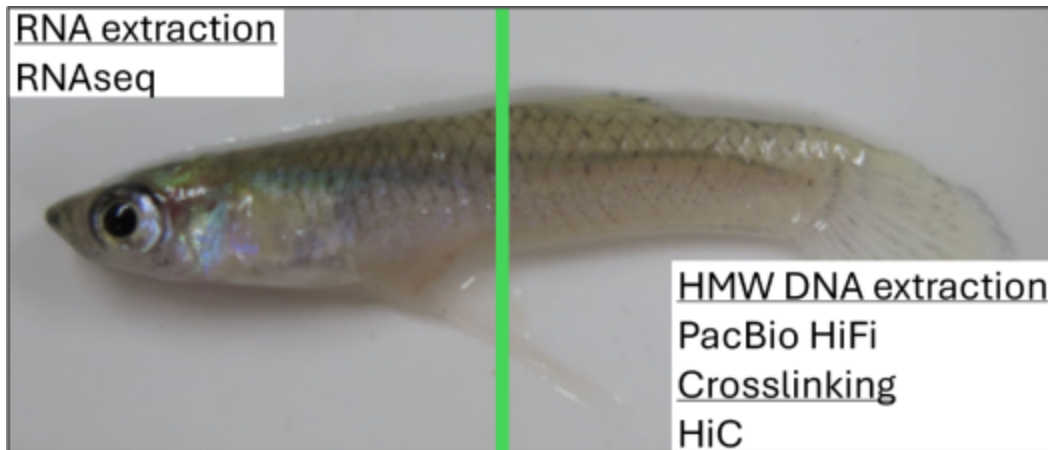
18. Chan, P. P., et al. (2021). "tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes." Nucleic acids research **49**(16): 9077-9096.
19. Chen, T., et al. (2010). "The role of female dominance hierarchies in the mating behaviour of mosquitofish." Biology Letters **7**(3): 343.
20. Cheng, H., et al. (2021). "Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm." Nature methods **18**(2): 170-175.
21. Condon, C. H. and R. S. Wilson (2006). "Effect of thermal acclimation on female resistance to forced matings in the eastern mosquitofish." Animal Behaviour **72**(3): 585-593.
22. Danecek, P., et al. (2021). "Twelve years of SAMtools and BCFtools." Gigascience **10**(2): giab008.
23. Di Tommaso, P., et al. (2017). "Nextflow enables reproducible computational workflows." Nature biotechnology **35**(4): 316-319.
24. Dobin, A., et al. (2013). "STAR: ultrafast universal RNA-seq aligner." Bioinformatics **29**(1): 15-21.
25. Durand, N. C., et al. (2016). "Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments." Cell systems **3**(1): 95-98.
26. Fraser, B., et al. (2026). "ERGA-BGE reference genome of *Gambusia holbrooki*, a globally invasive freshwater fish." Open Research Europe **6**(48): 48.
27. Frith, M. C. and R. Kawaguchi (2015). "Split-alignment of genomes finds orthologies more accurately." Genome biology **16**(1): 106.
28. Griffiths-Jones, S., et al. (2003). "Rfam: an RNA family database." Nucleic acids research **31**(1): 439-441.
29. Gurevich, A., et al. (2013). "QUAST: quality assessment tool for genome assemblies." Bioinformatics **29**(8): 1072-1075.
30. Harrison, L. M., et al. (2023). "The effect of brief or prolonged bouts of winning or losing male-male contests on plasticity in sexually selected traits." The American Naturalist **201**(3): 442-459.
31. Haynes, J. L. (1993). "Annual reestablishment of mosquitofish populations in Nebraska." Copeia **1993**(1): 232-235.
32. Huerta-Cepas, J., et al. (2019). "eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses." Nucleic acids research **47**(D1): D309-D314.
33. Iglesias-Carrasco, M., et al. (2019). "An experimental test for body size-dependent effects of male harassment and an elevated copulation rate on female lifetime fecundity and offspring performance." Journal of Evolutionary Biology **32**(11): 1262-1273.
34. Jawad, L. and A. Busneina (2000). "Fecundity of mosquitofish, *Gambusia affinis* (Baird & Girard) as a function of female size in fish from two lakes in Libya." Miscel·lània Zoològica: 31-40.
35. Kahn, A. T., et al. (2012). "Do females preferentially associate with males given a better start in life?" Biology Letters **8**(3): 362.

36. Kahn, A. T., et al. (2009). "Females prefer to associate with males with longer intromittent organs in mosquitofish." Biology Letters **6**(1): 55.
37. Kielbasa, S. M., et al. (2011). "Adaptive seeds tame genomic sequence comparison." Genome Research **21**(3): 487.
38. Kim, B., et al. (2021). "Male size and reproductive performance in three species of livebearing fishes (*Gambusia* spp.): a systematic review and meta-analysis." Journal of Animal Ecology **90**(10): 2431-2445.
39. Krumholz, L. A. (1948). "Reproduction in the western mosquitofish, *Gambusia affinis affinis* (Baird & Girard), and its use in mosquito control." Ecological Monographs **18**(1): 1-43.
40. Kurtul, I., et al. (2024). "Exploring invasiveness and versatility of used microhabitats of the globally invasive *Gambusia holbrooki*." Science of the Total Environment **925**: 171718.
41. Kurtzer, G. M., et al. (2017). "Singularity: Scientific containers for mobility of compute." PloS one **12**(5): e0177459.
42. Li, H. (2023). "Protein-to-genome alignment with miniprot." Bioinformatics **39**(1): btad014.
43. Lloyd, L. and J. Tomasov (1985). "Taxonomic status of the mosquitofish, *Gambusia affinis* (Poeciliidae), in Australia." Marine and Freshwater Research **36**(3): 447-451.
44. Loughland, I. and F. Seebacher (2020). "Differences in oxidative status explain variation in thermal acclimation capacity between individual mosquitofish (*Gambusia holbrooki*)." Functional Ecology **34**(7): 1380-1390.
45. Manni, M., et al. (2021). "BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes." Molecular biology and evolution **38**(10): 4647-4654.
46. Mapleson, D., et al. (2017). "KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies." Bioinformatics **33**(4): 574-576.
47. Matthews, S. A. and M. Y. Wong (2015). "Temperature-dependent resolution of conflict over rank within a size-based dominance hierarchy." Behavioral Ecology **26**(3): 947-958.
48. Meffe, G. K. (1990). "Offspring size variation in eastern mosquitofish (*Gambusia holbrooki*: Poeciliidae) from contrasting thermal environments." Copeia: 10-18.
49. Meffe, G. K. (1991). "Life history changes in eastern mosquitofish (*Gambusia holbrooki*) induced by thermal elevation." Canadian Journal of Fisheries and Aquatic Sciences **48**(1): 60-66.
50. Meffe, G. K. (1992). "Plasticity of life-history characters in eastern mosquitofish (*Gambusia holbrooki*: Poeciliidae) in response to thermal stress." Copeia: 94-102.
51. Mills, M. D., et al. (2004). "Complex interactions between native and invasive fish: the simultaneous effects of multiple negative interactions." Oecologia **141**(4): 713-721.
52. Nawrocki, E. and S. E. Infernal "1.1: 100-fold faster RNA homology searches., 2013, 29." DOI: <https://doi.org/10.1093/bioinformatics/btt509>: 2933-2935.
53. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRP690386> (2017). [See note]

54. O'dea, R., et al. (2014). "Male body size and condition affects sperm number and production rates in mosquitofish, *Gambusia holbrooki*." Journal of Evolutionary Biology **27**(12): 2739-2744.
55. Pilastro, A., et al. (2003). "Female aggregation and male competition reduce costs of sexual harassment in the mosquitofish *Gambusia holbrooki*." Animal Behaviour **65**(6): 1161-1167.
56. Plessy, C., et al. (2024). "Extreme genome scrambling in marine planktonic *Oikopleura dioica* cryptic species." Genome Research **34**(3): 426.
57. Pyke, G. H. (2008). "Plague minnow or mosquito fish? A review of the biology and impacts of introduced *Gambusia* species." Annual review of ecology, evolution, and systematics **39**(1): 171-191.
58. Rees, B. (1958). Attributes of the mosquito fish in relation to mosquito control. Proc. Calif. Mosq. Contr. Assoc.
59. Rhie, A., et al. (2020). "Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies." Genome biology **21**(1): 245.
60. Shao, F., et al. (2020). "Chromosome-level genome assembly of the female western mosquitofish (*Gambusia affinis*)." Gigascience **9**(8): giaa092.
61. Shen, W., et al. (2016). "SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation." PloS one **11**(10): e0163962.
62. Specziár, A. (2004). "Life history pattern and feeding ecology of the introduced eastern mosquitofish, *Gambusia holbrooki*, in a thermal spa under temperate climate, of Lake Hévíz, Hungary." Hydrobiologia **522**(1): 249-260.
63. Stoffels, R. J., et al. (2017). "Heat and hypoxia give a global invader, *Gambusia holbrooki*, the edge over a threatened endemic fish on Australian floodplains." Biological Invasions **19**(8): 2477-2489.
64. Storer, J., et al. (2021). "The Dfam community resource of transposable element families, sequence models, and genome annotations." Mobile DNA **12**(1): 2.
65. Uliano-Silva, M., et al. (2023). "MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads." BMC bioinformatics **24**(1): 288.
66. Vasimuddin, M., et al. (2019). Efficient architecture-aware acceleration of BWA-MEM for multicore systems. 2019 IEEE international parallel and distributed processing symposium (IPDPS), IEEE.
67. Walton, W. E., et al. (2012). "*Gambusia affinis* (Baird and Girard) and *Gambusia holbrooki* Girard (Mosquitofish)." A handbook of global freshwater invasive species: 261-273.
68. Wilson, R. S. (2005). "Temperature influences the coercive mating and swimming performance of male eastern mosquitofish." Animal Behaviour **70**(6): 1387-1394.
69. Wilson, R. S. and I. A. Johnston (2004). Combining studies of comparative physiology and behavioural ecology to test the adaptive benefits of thermal acclimation. International Congress Series, Elsevier.
70. Wong, B. B. and M. McCarthy (2009). "Prudent male mate choice under perceived sperm competition risk in the eastern mosquito fish." Behavioral Ecology **20**(2): 278-282.

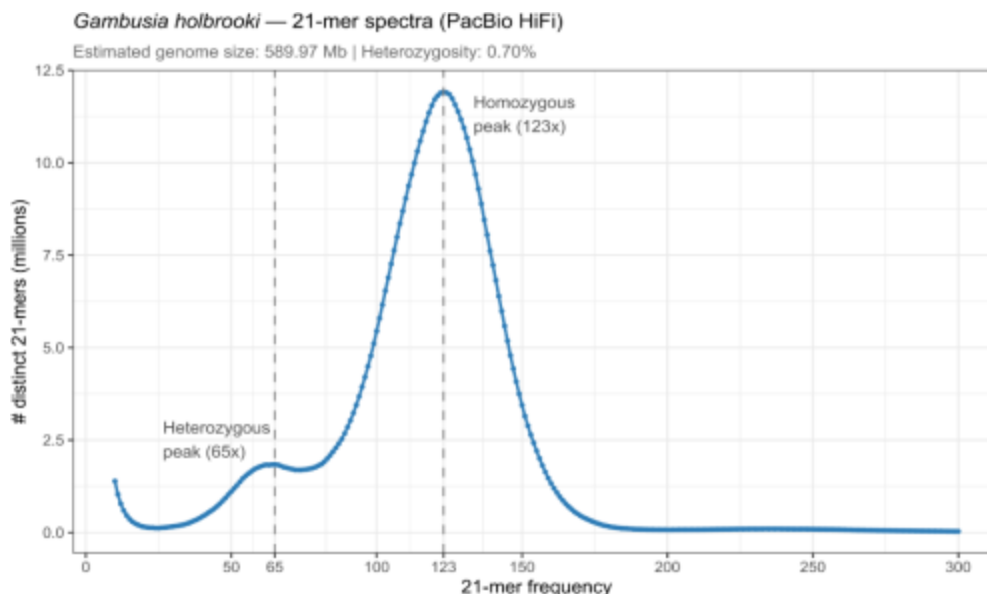
71. Zane, L., et al. (1999). "Microsatellite assessment of multiple paternity in natural populations of a live-bearing fish, *Gambusia holbrooki*." *Journal of Evolutionary Biology* **12**(1): 61-69.
72. Zhou, C., et al. (2023). "YaHS: yet another Hi-C scaffolding tool." *Bioinformatics* **39**(1): btac808.

## Figures



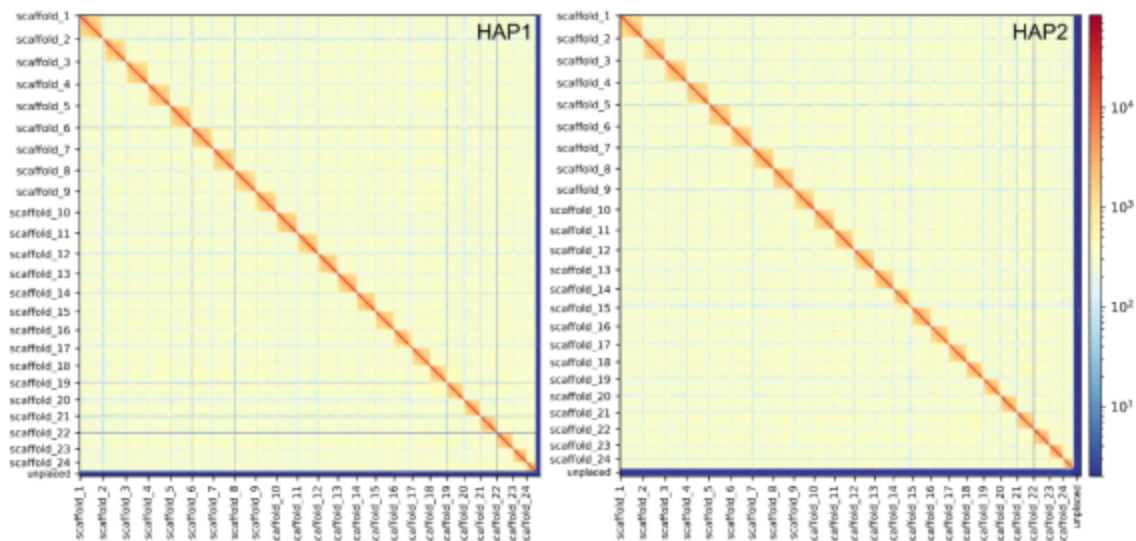
**Figure 1**

Sampling of male mosquitofish. Body before anal pore was used for RNA extraction, while the tail muscle was used for HMW DNA extraction for PacBio HiFi sequencing and Chromatin Crosslinking for HiC sequencing.



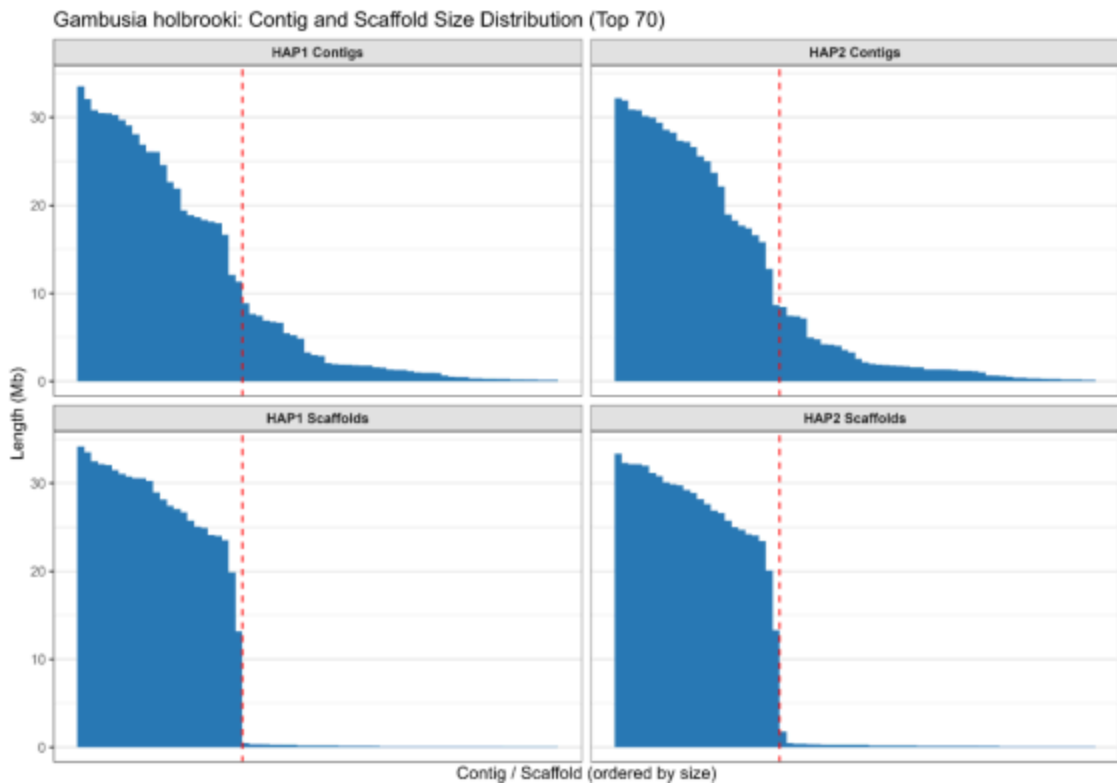
**Figure 2**

PacBio HiFi reads 21-mer frequency spectra. The heterozygous peak (~68x) and homozygous peak (~123x) are consistent with a diploid genome at ~123x haploid coverage and 0.70% heterozygosity.



**Figure 3**

Hi-C contact maps for the *Gambusia holbrooki* HAP1 (left) and HAP2 (right) assemblies. Contact frequency is shown on a log10 scale (blue to red). Strong diagonal signal across all 24 chromosome-scale scaffolds confirms high-quality chromosomal assignment, with minimal inter-chromosomal contacts and a small unplaced fraction in both haplotypes.



**Figure 4**

Contig and scaffold size distribution (top 70) for HAP1 and HAP2. The red dashed line marks position 24, corresponding to the expected number of chromosomes. Left of the line: chromosome-scale

sequences. Right of the line: unplaced contigs/scaffolds.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementalFigures.docx](#)