DATA NOTE

REVISED ERGA-BGE reference genome of *Gambusia holbrooki*, a globally invasive freshwater fish

[version 2; peer review: 2 approved]

Bonnie Fraser[1], Clelia Gasparini[2], Francesco Santi[2], Astrid Böhne[3], Rita Monteiro[3], Thomas Marcussen[4,5], Rebekah A. Oomen[4-8], Torsten H. Struck[4],
Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory Team,
Wellcome Sanger Institute Scientific Operations: Sequencing Operations,
Wellcome Sanger Institute Tree of Life Core Informatics Team, Amy Denton[9], Caroline Howard[9], Kerstin Howe[9], Mark Blaxter[9], Shane McCarthy[9], Jonathan M.D. Wood[9], Fergal Martin[10], Swati Sinha[10], Leanne Haggerty[10], Chiara Bortoluzzi[11]

[1]Biosciences, University of Exeter, Exeter, EX4 4SB, UK
[2]Department of Biology, Via U. Bassi, University of Padova, Padova, 58/B, Italy
[3]Leibniz Institute for the Analysis of Biodiversity Change, Museum Koenig Bonn, Bonn, 53113, Germany
[4]Natural History Museum, P.O. Box 1172, Blindern, University of Oslo, Oslo, 0318, Norway
[5]Centre for Ecological & Evolutionary Synthesis, University of Oslo, Oslo, Norway
[6]Department of Biological Sciences, University of New Brunswick Saint John, Saint John, Canada
[7]Tjärnö Marine Laboratory, University of Gothenburg, Gothenburg, Sweden
[8]Centre for Coastal Research, University of Agder, Kristiansand, Norway
[9]Tree of Life, Wellcome Sanger Institute, Hinxton, UK
[10]Wellcome Genome Campus, European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, UK
[11]SIB Swiss Institute of Bioinformatics, Amphipôle, Quartier UNIL-Sorge, Lausanne, 1015, Switzerland

v2 **First published:** 13 Feb 2026, **6**:48
https://doi.org/10.12688/openreseurope.22693.1
Latest published: 04 Jun 2026, **6**:48
https://doi.org/10.12688/openreseurope.22693.2

Abstract

The *Gambusia holbrooki* (eastern mosquitofish) reference genome will offer a crucial resource for understanding the evolution and adaptation of invasive freshwater fish species. The genome of *G. holbrooki* was assembled into two haplotypes through a phased assembly approach; however, only the primary haplotype was designated as the reference genome for annotation and downstream analyses. The entirety of the genome sequence was assembled into 24 contiguous chromosomal pseudomolecules and 1 mitochondrial genome. This chromosome-level assembly encompasses 0.67 Gb,

Open Peer Review

Approval Status ✓ ✓

	1	2
version 2 (revision) 04 Jun 2026	✓ view	✓ view
version 1 13 Feb 2026	? view	? view

1. **Alejandro S. Mechaly**, Instituto de

composed of 421 contigs and 318 scaffolds, with contig and scaffold N50 values of 15.9 Mb and 29.6 Mb, respectively.

Keywords
Gambusia holbrooki, genome assembly, European Reference Genome Atlas, Biodiversity Genomics Europe, Earth Biogenome Project, Eastern mosquitofish, Poeciliidae, invasive alien species

This article is included in the Horizon Europe gateway.

This article is included in the Genome Reports from the Biodiversity Genomics Europe Project collection.

Investigaciones en Biodiversidad y Biotecnología (INBIOTEC-CONICET), Mar del Plata, Argentina

2. **Roger Huerlimann**, Okinawa Institute of Science and Technology, Okinawa, Japan

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

Corresponding author: Chiara Bortoluzzi (chiara.bortoluzzi@sib.swiss)

Author roles: Fraser B: Investigation, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Gasparini C**: Investigation, Resources, Writing – Review & Editing; **Santi F**: Investigation, Resources, Writing – Review & Editing; **Böhne A**: Methodology, Project Administration, Supervision, Writing – Review & Editing; **Monteiro R**: Methodology, Project Administration, Supervision, Writing – Review & Editing; **Marcussen T**: Methodology, Project Administration, Supervision, Writing – Review & Editing; **Oomen RA**: Methodology, Project Administration, Supervision, Writing – Review & Editing; **Struck TH**: Methodology, Project Administration, Supervision, Writing – Review & Editing; **Denton A**: Investigation, Writing – Review & Editing; **Howard C**: Investigation, Writing – Review & Editing; **Howe K**: Data Curation, Formal Analysis, Supervision, Writing – Review & Editing; **Blaxter M**: Data Curation, Formal Analysis, Supervision, Writing – Review & Editing; **McCarthy S**: Data Curation, Formal Analysis, Writing – Review & Editing; **Wood JMD**: Data Curation, Formal Analysis, Writing – Review & Editing; **Martin F**: Data Curation, Formal Analysis, Writing – Review & Editing; **Sinha S**: Data Curation, Formal Analysis, Writing – Review & Editing; **Haggerty L**: Data Curation, Formal Analysis, Writing – Review & Editing; **Bortoluzzi C**: Visualization, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: Biodiversity Genomics Europe (Grant no. 101059492) is funded by Horizon Europe under the Biodiversity, Circular Economy and Environment call (REA.B.3); co-funded by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract numbers 22.00173 and 24.00054; and by the UK Research and Innovation (UKRI) under the Department for Business, Energy and Industrial Strategy's Horizon Europe Guarantee Scheme.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Fraser B, Gasparini C, Santi F *et al.* **ERGA-BGE reference genome of *Gambusia holbrooki*, a globally invasive freshwater fish [version 2; peer review: 2 approved]** Open Research Europe 2026, **6**:48
https://doi.org/10.12688/openreseurope.22693.2

First published: 13 Feb 2026, **6**:48 https://doi.org/10.12688/openreseurope.22693.1

REVISED **Amendments from Version 1**

In this revised version, we addressed all reviewers comments. In particular, we clarified the reason behind the low BUSCO completeness score of the genome annotation, as pointed out by both reviewers. We also provided more context in the introduction, by clarifying the relevance of this genome to the comparative genomics community as it adds up to the growing number of genomes from the Poeciliidae family.

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

Introduction

Gambusia holbrooki, a member of the *Poeciliidae* family, is a freshwater fish native to eastern North America (Pyke, 2005). Along with its sister species *Gambusia affinis*, it has been introduced globally as a means of mosquito control, with often severe impacts on local biodiversity (Fryxell *et al.*, 2022). In Europe, *Gambusia holbrooki* was introduced to Spain and Italy in 1921 and has since become established throughout the Southern part of the European continent (Fryxell *et al.*, 2022). Both species have been introduced into Argentina (Cabrera *et al.*, 2017) and *G. holbrooki* have been identified in Australia (Lloyd & Tomasov, 1985). However, taxonomic confusion has been documented. Mosquitofish (*G. holbrooki* and *G. affinis*) can have devastating impacts on local biodiversity and has been listed among the top 100 invasive alien species (Lowe *et al.*, 2000). Being omnivorous and tolerant to a variety of environmental conditions, mosquitofish quickly spread with negative consequences for local biodiversity. In addition to its well-documented invasion history, the mosquitofish is a longstanding model system for life history evolution research and was key in the development of life history theory (Stearns, 1983). Finally, *G. holbrooki* is a member of the Poeciliidae family, a rapidly growing model system for comparative genomics (e.g., Ryan *et al.*, 2025). *Gambusia holbrooki* is listed as Least Concern (LC) in the IUCN Red List on Threatened Species (last assessed in 2012).

The generation of this reference resource was coordinated by the European Reference Genome Atlas (ERGA) initiative's Biodiversity Genomics Europe (BGE) project, supporting ERGA's aims of promoting transnational cooperation to promote advances in the application of genomics technologies to protect and restore biodiversity (Mazzoni *et al.*, 2023).

Materials & methods

ERGA's sequencing strategy includes Pacific Biosciences (PacBio) for long-read sequencing, along with Hi-C sequencing for chromosomal architecture, and RNA sequencing for transcriptomic profiling, to facilitate genome assembly and annotation.

Sample and sampling information

On 24 July 2024, Clelia Gasparini and Francesco Santi sampled 8 male specimens of *Gambusia holbrooki*, which were identified based on morphology using the gonopodium structure (Lloyd & Tomasov, 1985) at the University of Padua, Italy. Specimens were sampled in the Piovego canal (which connects the Bacchiglione and Brenta rivers), Padua, Italy. *Gambusia holbrooki* is the only mosquitofish species that has so far been described in Italy. Sampling was performed under permission number 382 (16 May 2024) issued by the Regione del Veneto, Italy. Specimens were caught with a dip net and immediately placed into a collection container. Specimens were then euthanized with an overdose of MS222 and subsequently put in dry ice. Until DNA extraction, samples were preserved at −80 °C.

Vouchering information

Physical reference materials for the here sequenced specimen have been deposited in the Leibniz Institute for the Analysis of Biodiversity Change https://leibniz-lib.de/en/index.html under accession numbers ZFMK-TIS-97651 (male organism) and ZFMK-TIS-97652 (female organism).

Frozen somatic animal tissue is available from proxy individuals at the Biobank of the Leibniz Institute for the Analysis of Biodiversity Change https://leibniz-lib.de/en/index.html under the voucher IDs ZFMK-TIS-97677, ZFMK-TIS-97679, ZFMK-TIS-97680, ZFMK-TIS-97682 to ZFMK-TIS-97685, ZFMK-TIS-97688, ZFMK-TIS-97689, ZFMK-TIS-97692. Additionally, somatic cell cultures were established and frozen in liquid nitrogen under the ID ZFMK-TIS-98713.

Genetic information

The estimated genome size, based on ancestral taxa, is 0.83 Gb. This is a diploid genome with a haploid number of 24 chromosomes (2n = 48) and XX/XY as sex determination mechanism (Kottler *et al.*, 2020). All information for this species was retrieved from Genomes on a Tree (Challis *et al.*, 2023).

DNA/RNA processing

Protocols for high molecular weight (HMW) DNA extraction developed at the Wellcome Sanger Institute (WSI) Tree of Life Core Laboratory are available on protocols.io (Denton *et al.*, 2023b; Howard *et al.*, 2025). The fGamHol4 sample was weighed and triaged (Jay *et al.*, 2023) to determine the appropriate extraction protocol. Tissue from the mid-body was homogenised by powermashing using a PowerMasher II tissue disruptor (Denton *et al.,* 2023a). HMW DNA was extracted using the Automated MagAttract v2 protocol (Oatley *et al.,* 2025a). DNA was sheared into an average fragment size of 12–20 kb following the Megaruptor®3 for LI PacBio protocol (Bates *et al.,* 2023). Sheared DNA was purified by automated SPRI (solid-phase reversible immobilisation) (Oatley *et al.*, 2025b). The concentration of the sheared and purified DNA was assessed using a Nanodrop spectrophotometer and Qubit Fluorometer using the Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the FemtoPulse system.

Library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core. Libraries were prepared using the SMRTbell Prep Kit 3.0 (Pacific Biosciences, California, USA), according to the manufacturer's instructions. The kit includes reagents for end repair/A-tailing, adapter ligation, post-ligation SMRTbell bead clean-up, and nuclease treatment. Size selection and clean-up were performed using diluted AMPure PB beads (Pacific Biosciences). DNA concentration was quantified using a Qubit Fluorometer v4.0 (ThermoFisher Scientific) and the Qubit 1X dsDNA HS assay kit. Final library fragment size was assessed with the Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies) using the gDNA 55 kb BAC analysis kit.

The sample was sequenced on a Revio instrument (Pacific Biosciences). The prepared library was normalised to 2 nM, and 15 µL was used for making complexes. Primers were annealed and polymerases bound to generate circularised complexes, following the manufacturer's instructions. Complexes were purified using 1.2X SMRTbell beads, then diluted to the Revio loading concentration (200–300 pM) and spiked with a Revio sequencing internal control. The sample was sequenced on a Revio 25 M SMRT cell. The SMRT Link software (Pacific Biosciences), a web-based workflow manager, was used to configure and monitor the run and to carry out primary and secondary data analysis.

Biotinylated DNA constructs were fragmented using a Covaris E220 sonicator and size selected to 400–600 bp using SPRISelect beads. DNA was enriched with Arima-HiC v2 kit Enrichment beads. End repair, A-tailing, and adapter ligation were carried out with the NEBNext Ultra II DNA Library Prep Kit (New England Biolabs), following a modified protocol where library preparation occurs while DNA remains bound to the Enrichment beads. Library amplification was performed using KAPA HiFi HotStart mix and a custom Unique Dual Index (UDI) barcode set (Integrated DNA Technologies). Depending on sample concentration and biotinylation percentage determined at the crosslinking stage, libraries were amplified with 10–16 PCR cycles. Post-PCR clean-up was performed with SPRISelect beads. Libraries were quantified using the AccuClear Ultra High Sensitivity dsDNA Standards Assay Kit (Biotium) and a FLUOstar Omega plate reader (BMG Labtech). Prior to sequencing, libraries were normalised to 10 ng/µL. Normalised libraries were quantified again and equimolar and/or weighted 2.8 nM pools. Pool concentrations were checked using the Agilent 4200 TapeStation (Agilent) with High Sensitivity D500 reagents before sequencing. Sequencing was performed using paired-end 150 bp reads on the Illumina NovaSeq X. In total 56x HiFi and 147x HiC data were sequenced to generate the assembly.

Genome assembly methods

The HiFi reads were assembled using Hifiasm (Cheng *et al.*, 2021) in Hi-C phasing mode, where data were separated into two haplotypes. These haplotypes were then curated to generate a final assembly. The Hi-C reads were aligned to the contigs using bwa-mem2 (Vasimuddin *et al.*, 2019), and contigs were scaffolded with YaHS (Zhou *et al.*, 2023), using the --break option for handling potential misassemblies. The resulting scaffolded assemblies were evaluated using Gfastats (Formenti *et al.*, 2022), BUSCO (Manni *et al.*, 2021), and MERQURY.FK (Rhie *et al.*, 2020).

The mitochondrial genome was assembled using oatk (Zhou *et al.*, 2024) as a single circular contig of 18,667 bp and it is included in the released assembly (GCA_965178065.1). Manual curation was primarily conducted using PretextView (Harry, 2022), with additional insights provided by JBrowse2 (Diesh *et al.*, 2023) and HiGlass (Kerpedjiev *et al.*, 2018). Scaffolds were visually inspected and corrected as described by (Howe *et al.*, 2021). Any identified contamination, missed joins, and mis-joins were corrected, and duplicate sequences were tagged and removed. The curation process is documented at (Denton *et al.*, 2023) (article in preparation). Summary analysis of the released assembly was performed using the ERGA-BGE Genome Report ASM Galaxy workflow (doi.org/10.48546/workflowhub.workflow.1104.1).

Genome annotation methods

A gene set was generated using the Ensembl Gene Annotation system (Aken *et al.*, 2016), primarily by aligning publicly available short-read RNA-seq data from BioSample: SAMN14167175, SAMN14167173, SAMN11245730, SAMN11245729, SAMN11245733, SAMN11245732, SAMN14167176, SAMN14167159, SAMN14167158, SAMN11245734, SAMN11245731, SAMN14167178, SAMN14167177, SAMN14167174, SAMN14167160, SAMN14167157, SAMN14167156, and SAMN14167155 to the genome. Gaps in the annotation were filled via protein-to-genome alignments of a select set of vertebrate proteins from (UniProt Consortium, 2019), which had experimental evidence at the protein or transcript level. At each locus, data were aggregated and consolidated, prioritising models derived from RNA-seq data, resulting in a final set of gene models and associated non-redundant transcript sets. To distinguish true isoforms from fragments, the likelihood of each open reading frame (ORF) was evaluated against known vertebrate proteins. Low-quality transcript models, such as those showing evidence of fragmented ORFs, were removed. In cases where RNA-seq data were fragmented or absent, homology data were prioritised, favouring longer transcripts with strong intron support from short-read data. The resulting gene models were classified into three categories: protein-coding, pseudogene, and long non-coding. Models with hits to known proteins and few structural abnormalities were classified as protein-coding. Models with hits to known proteins but displaying abnormalities, such as the absence of a start codon, non-canonical splicing, unusually small intron structures (<75 bp), or excessive repeat coverage, were reclassified as pseudogenes. Single-exon models with a corresponding multi-exon copy elsewhere in the genome were classified as processed (retrotransposed) pseudogenes. Models that did not fit any of the previously

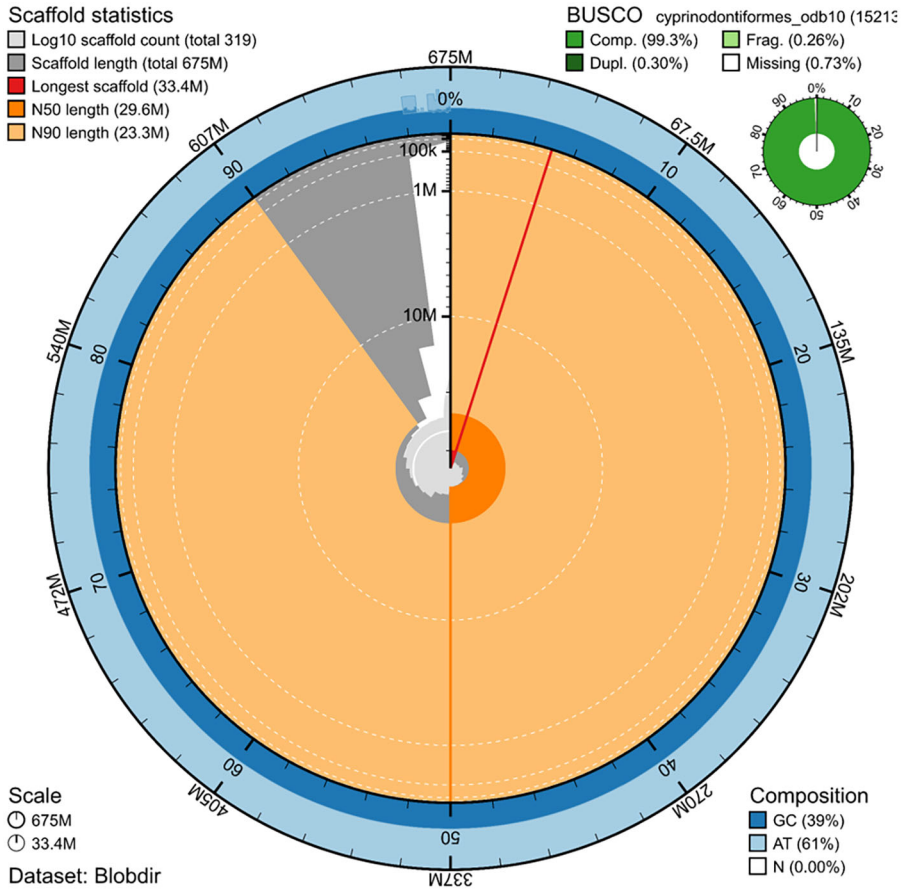

Figure 1. Snail plot summary of assembly statistics. The main plot is divided into 1,000 size-ordered bins around the circumference, with each bin representing 0.1% of the 674,622,098 bp assembly including the mitochondrial genome. The distribution of sequence lengths is shown in dark grey, with the plot radius scaled to the longest sequence present in the assembly (33.4 Mp, shown in red). Orange and pale-orange arcs show the scaffold N50 and N90 sequence lengths (29,628,722 and 23,341,799 bp), respectively. The pale grey spiral shows the cumulative sequence count on a log-scale, with white scale lines showing successive orders of magnitude. The blue and pale-blue area around the outside of the plot shows the distribution of GC, AT, and N percentages in the same bins as the inner plot. A summary of complete, fragmented, duplicated, and missing BUSCO genes found in the assembled genome from the Cyprinodontiformes database (odb10) is shown in the top right.

described categories did not overlap protein-coding genes and were constructed from transcriptomic data were considered potential lncRNAs. Potential lncRNAs were further filtered to remove single-exon loci due to their unreliability. Putative miRNAs were predicted by performing a BLAST search of miRBase (Kozomara *et al.*, 2019) against the genome, followed by RNAfold analysis (Gruber *et al.*, 2008). Other small non-coding loci were identified by scanning the genome with Rfam (Kalvari *et al.*, 2018) and passing the results through Infernal (Nawrocki & Eddy, 2013). Summary analysis of the released annotation was carried out using the ERGA-BGE Genome Report ANNOT Galaxy workflow (10.48546/workflowhub.workflow.1096.1).

Results

Genome assembly

The genome assembly has a total length of 674,622,098 bp in 318 scaffolds including the mitogenome (Figure 1 & Figure 2), with a GC content of 39%. The assembly has a contig N50 of 15,918,108 bp and L50 of 16 and a scaffold N50 of 29,628,722 bp and L50 of 11. The assembly has a total of 103 gaps, totalling 11 kb in cumulative size. The single-copy gene content analysis using the Cyprinodontiformes database (odb10) with BUSCO (Manni *et al.*, 2021) resulted in 99.3% completeness (99.0% single and 0.3% duplicated). 93.8% of reads k-mers were present in the assembly and the assembly has a base accuracy Quality Value (QV) of 59.8 as calculated by Merqury (Rhie *et al.*, 2020).

Genome annotation

The genome annotation consists of 21,050 protein-coding genes with associated 56,794 transcripts, in addition to 3,687 non-coding genes (Table 1). Using the longest isoform per transcript, the single-copy gene content analysis using the Cyprinodontiformes database (odb10) database with BUSCO resulted in 88.5% completeness. The discrepancy between the very high assembly-level BUSCO completeness (99.3%) and the lower protein-based BUSCO score for the annotation likely reflects characteristics of the automated annotation workflow. The annotation was generated using

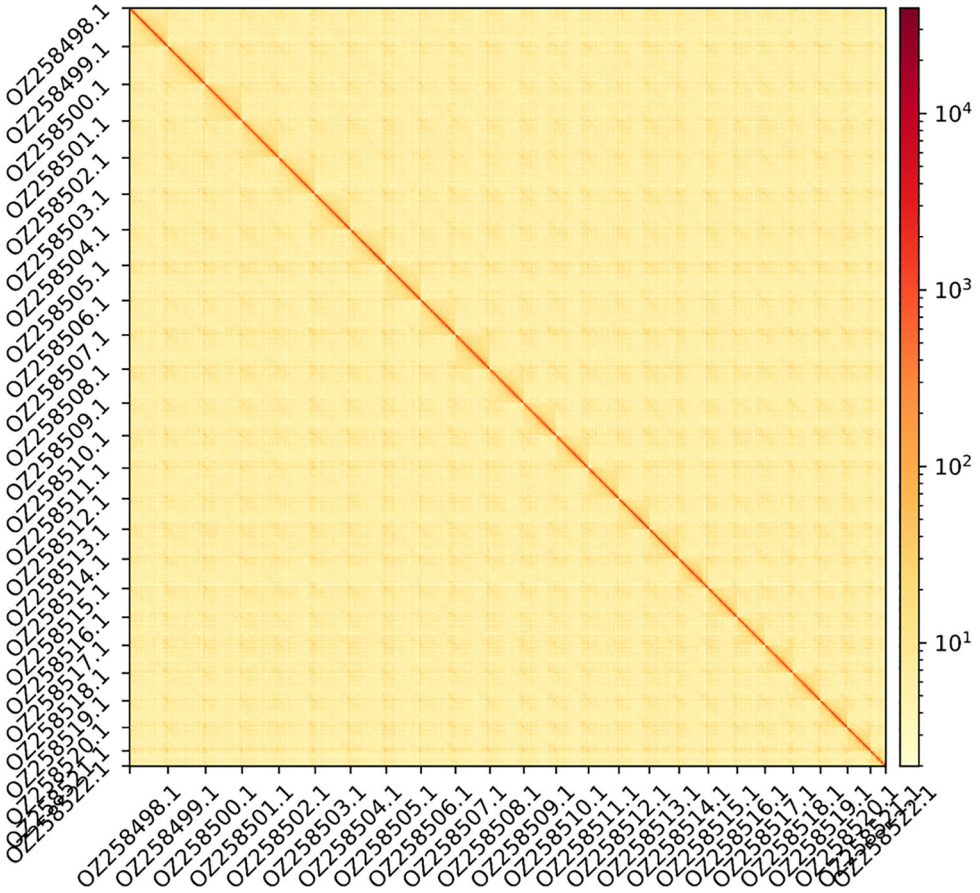

Figure 2. Hi-C contact map showing spatial interactions between regions of the genome. The diagonal corresponds to intra-chromosomal contacts, depicting chromosome boundaries. The frequency of contacts is shown on a logarithmic heatmap scale. Hi-C matrix bins were merged into a 200 kb bin size for plotting.

Table 1. Statistics from assembled gene models.

	No. genes	No. transcripts	Mean gene length (bp)	No. single-exon genes	Mean exons per transcript
mRNA	21,050	56,794	18,920	517	14.3
pseudogene	228	228	22,143	1	24.5
snoRNA	158	158	114	158	1.0
lncRNA	2,085	2,520	4,158	1,283	2.6
miRNA	25	25	69	25	1.0
snRNA	118	118	153	118	1.0
rRNA	1,066	1,066	612	1,066	1.0
scRNA	7	7	223	7	1.0
Other ncRNA	13	13	98–275	13	1.0–1.0

Table 2. Annotation completeness and consistency scores calculated by BUSCO run in protein mode (cyprinodontiformes_odb10) and OMArk (Metazoa-v2.0.0.h5).

	Complete	Singular	Duplicated	Fragmented	Missing
BUSCO	13,470 (88.5%)	13,391 (88.0%)	79 (0.5%)	197 (1.3%)	1,546 (10.2%)
OMArk	17,336 (94.0%)	17,053 (92.5%)	283 (1.5%)	-	1,105 (5.6%)
	Consistent	**Inconsistent**	**Contaminants**	**Unknown**	
OMArk	20,764 (98.5%)	142 (0.7%)	0.0 (0.0%)	166 (0.8%)	

the automated Ensembl annotation pipeline, which applies relatively stringent filtering criteria during gene model construction. In particular, transcript-supported models are prioritised when they are additionally supported by protein homology evidence, including similarity searches against curated UniProt protein datasets. Such conservative filtering can reduce false-positive annotations but may also exclude partial, lowly expressed, lineage-specific, or weakly supported gene models, thereby lowering annotation BUSCO completeness relative to the assembly BUSCO score. In addition, the annotation relied primarily on short-read RNA-seq data, which may not fully capture transcript diversity, splice isoforms, or genes expressed only in specific tissues or developmental stages. Using the OMAmer Metazoa-v2.0.0.h5 database for OMArk (Nevers *et al.*, 2025), the genome annotation has a completeness score of 94.0% and a consistency score of 98.5% (Table 2).

Data availability

Gambusia holbrooki and the related genomic study were assigned to Tree of Life ID (ToLID) 'fGamHol4' and all sample, sequence, and assembly information are available under the umbrella BioProject PRJEB84145. The sample information is available at the following BioSample accessions: SAMEA115931355 and SAMEA115931357. The genome assembly is accessible from ENA under accession number GCA_965282415.1 and the annotated genome is available through the Ensembl website (https://projects.ensembl.org/erga-bge/). Sequencing data produced as part of this project are available from ENA at the following accessions: ERX13553727 and ERX14241339. Documentation related to the genome assembly and curation can be found in the ERGA Assembly Report (EAR) document available at https://github.com/ERGA-consortium/EARs/tree/main/Assembly_Reports/Gambusia_holbrooki/fGamHol4. Further details and data about the project are hosted on the ERGA portal at https://portal.erga-biodiversity.eu/data_portal/37273.

Author contributions

BF and CG Fraser coordinated the project; CG and FS collected the species; FS identified the species; FS and CG sampled and preserved biological material and provided metadata; AsB, RM, TM, RAO, and THS provided support in sampling, shipping of biological material, metadata collection, and management; DNA extraction, library preparation and sequencing were performed by AD, CH and the teams at the Wellcome Sanger Institute; assembly and curation was carried out by SMC and JDMW under the supervision of KH and MB; FM, SS, and LH performed genome annotation; CB generated the analysis and report. All authors contributed to the writing, review, and editing of this genome note and read and approved the final version.

Acknowledgements

We acknowledge the support of the Freiburg Galaxy Team: Saim Momin and Björn Grüning, Bioinformatics, University of Freiburg (Germany), funded by the German Federal Ministry of Education and Research BMBF grant 031 A538A de. NBI-RBC and the Ministry of Science, Research and the Arts Baden-Württemberg (MWK) within the framework of LIBIS/de. NBI Freiburg. We would like to acknowledge the assembly reviewer, Jessica Gomez Garrido from the Centro Nacional de Análisis Genómico.

References

Aken BL, Ayling S, Barrell D, *et al.*: **The Ensembl gene annotation system.** *Database (Oxford).* 2016; **2016**: baw093.
PubMed Abstract | Publisher Full Text | Free Full Text

Bates A, Clayton-Lucey I, Howard C: **Sanger Tree of Life HMW DNA fragmentation: diagenode megaruptor® 3 for LI pacBio.** 2023.
Reference Source

Cabrera MB, Bogan S, Posadas P, *et al.*: **Risks associated with introduction of poeciliiids for control of mosquito larvae: first record of the non-native Gambusia holbrooki in Argentina.** *J Fish Biol.* 2017; **91**(2): 704–710.
PubMed Abstract | Publisher Full Text

Challis R, Kumar S, Sotero-Caio C, *et al.*: **Genomes on a Tree (GoaT): a versatile, scalable search engine for genomic and sequencing project metadata across the eukaryotic Tree of Life [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2023; **8**: 24.
PubMed Abstract | Publisher Full Text | Free Full Text

Cheng H, Concepcion GT, Feng X, *et al.*: **Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm.** *Nat Methods.* 2021; **18**(2): 170–175.
PubMed Abstract | Publisher Full Text | Free Full Text

Denton A, Oatley G, Cornwell C, *et al.*: **Sanger Tree of Life sample homogenisation: powerMash.** 2023a.
Publisher Full Text

Denton A, Yatsenko H, Jay J, *et al.*: **Sanger Tree of Life wet laboratory protocol collection V.1.** 2023b.
Publisher Full Text

Diesh C, Stevens GJ, Xie P, *et al.*: **JBrowse 2: a modular genome browser with views of synteny and structural variation.** *Genome Biol.* 2023; **24**(1): 74.
PubMed Abstract | Publisher Full Text | Free Full Text

Formenti G, Abueg L, Brajuka A, *et al.*: **Gfastats: conversion, evaluation and manipulation of genome sequences using assembly graphs.** *Bioinformatics.* 2022; **38**(17): 4214–4216.
PubMed Abstract | Publisher Full Text | Free Full Text

Fryxell DC, Moffett ER, Kinnison MT, *et al.*: **From southern swamps to cosmopolitan model: humanity's unfinished history with mosquitofish.** *Fish and Fisheries.* 2022; **23**(1): 143–161.
Publisher Full Text

Gruber AR, Lorenz R, Bernhart SH, *et al.*: **The vienna RNA websuite.** *Nucleic Acids Res.* 2008; **36**: W70–W74.
Publisher Full Text

Harry E: **PretextView (Paired REad TEXTure Viewer): a desktop application for viewing pretext contact maps.** 2022.
Reference Source

Howard C, Denton A, Jackson B, *et al.*: **On the path to reference genomes for all biodiversity: lessons learned and laboratory protocols created in the Sanger Tree of Life core laboratory over the first 2000 species.** *bioRxiv.* 2025: 2025–2004.
Publisher Full Text

Howe K, Chow W, Collins J, *et al.*: **Significantly improving the quality of genome assemblies through curation.** *GigaScience.* 2021; **10**(1): giaa153.
PubMed Abstract | Publisher Full Text | Free Full Text

Jay J, Yatsenko H, Narváez-Gómez JP, *et al.*: **Sanger Tree of Life sample preparation: triage and dissection.** 2023.
Reference Source

Kalvari I, Nawrocki EP, Argasinska J, *et al.*: **Non-coding RNA analysis using the Rfam database.** *Curr Protoc Bioinformatics.* 2018; **62**(1): e51.
PubMed Abstract | Publisher Full Text | Free Full Text

Kerpedjiev P, Abdennur N, Lekschas F, *et al.*: **HiGlass: Web-based visual exploration and analysis of genome interaction maps.** *Genome Biol.* 2018; **19**(1): 125.
PubMed Abstract | Publisher Full Text | Free Full Text

Kottler VA, Feron R, Nanda I, *et al.*: **Independent origin of XY and ZW sex determination mechanisms in mosquitofish sister species.** *Genetics.* 2020; **214**(1): 193–209.
PubMed Abstract | Publisher Full Text | Free Full Text

Kozomara A, Birgaoanu M: **Griffiths-Jones S: miRBase: from microRNA sequences to function.** *Nucleic Acids Res.* 2019; **47**(D1): D155–D162.
PubMed Abstract | Publisher Full Text | Free Full Text

Lloyd LN, Tomasov JF: **Taxonomic status of the mosquitofish, *Gambusia affinis* (Poeciliidae), in Australia.** *Mar Freshw Res.* 1985; **36**(3): 447–451.
Publisher Full Text

Lowe S, Browne M, Boudjelas S, *et al.*: 2000; *100 of the world's worst invasive alien species: A selection from the global invasive species database. (Vol. 12),* vol. **12**Auckland: Invasive Species Specialist Group.
Reference Source

Manni M, Berkeley MR, Seppey M, *et al.*: **BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes.** *Mol Biol Evol.* 2021; **38**(10): 4647–4654.
PubMed Abstract | Publisher Full Text | Free Full Text

Mazzoni CJ, Ciofi C, Waterhouse RM: **Biodiversity: an atlas of European reference genomes.** *Nature.* 2023; **619**: 252.
Publisher Full Text

Nawrocki EP, Eddy SR: **Infernal 1.1: 100-fold faster RNA homology searches.** *Bioinformatics.* 2013; **29**(22): 2933–2935.
PubMed Abstract | Publisher Full Text | Free Full Text

Nevers Y, Warwick Vesztrocy A, Rossier V, *et al.*: **Quality assessment of gene repertoire annotations with OMArk.** *Nat Biotechnol.* 2025; **43**(1): 124–133.
PubMed Abstract | Publisher Full Text | Free Full Text

Oatley G, Denton A, Kovalevskaia A, *et al.*: **Sanger Tree of Life HMW DNA extraction: automated magAttract v.3.** 2025a.
Publisher Full Text

Oatley G, Sampaio F, Denton A, *et al.*: **Sanger Tree of Life fragmented DNA clean up: automated SPRI.** 2025b.
Publisher Full Text

Pyke GH: **A review of the biology of *Gambusia affinis* and *G. holbrooki*. *Rev Fish Biol. Fisheries.*** 2005; **15**(4): 339–365.
Publisher Full Text

Rhie A, Walenz BP, Koren S, *et al.*: **Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies.** *Genome Biol.* 2020; **21**(1): 245.
PubMed Abstract | Publisher Full Text | Free Full Text

Ryan K, De-Kayne R, Davis J, *et al.*: **New genome assemblies for Poeciliidae: A foundation for adaptation studies.** *Genome Biol Evol.* 2025; **17**(6): evaf111.
PubMed Abstract | Publisher Full Text | Free Full Text

Stearns SC: **The genetic basis of differences in life-history traits among six populations of mosquitofish (*Gambusia affinis*) that shared ancestors in 1905.** *Evolution.* 1983; **37**(3): 618–627.
PubMed Abstract | Publisher Full Text

UniProt Consortium: **UniProt: a worldwide hub of protein knowledge.** *Nucleic Acids Res.* 2019; **47**(D1): D506–D515.
PubMed Abstract | Publisher Full Text | Free Full Text

Vasimuddin M, Misra S, Li H, *et al.*: **Efficient architecture-aware acceleration of BWA-MEM for multicore systems.** *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS).* 2019; 314–324.
Reference Source

Zhou C, Brown M, Blaxter M, *et al.*: **Oatk: a *de novo* assembly tool for complex plant organelle genomes.** *bioRxiv.* 2024; 2024–10.
Reference Source

Zhou C, McCarthy SA, Durbin R: **YaHS: Yet another Hi-C Scaffolding tool.** *Bioinformatics.* 2023; **39**(1): btac808.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 2

Reviewer Report 23 June 2026

https://doi.org/10.21956/openreseurope.26126.r75454

✔ **Roger Huerlimann**
Okinawa Institute of Science and Technology, Okinawa, Japan

The authors have addressed all of my concerns, and I approve the latest version for acceptance.

Competing Interests: No competing interests were disclosed.

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 12 June 2026

https://doi.org/10.21956/openreseurope.26126.r75453

✔ **Alejandro S. Mechaly**
Instituto de Investigaciones en Biodiversidad y Biotecnología (INBIOTEC-CONICET), Mar del Plata, Argentina

The manuscript is now suitable for indexing. The authors have adequately addressed my previous concerns and have implemented the suggested revisions in a satisfactory manner. I have no further major comments and recommend indexing of the manuscript in its current form.

Competing Interests: No competing interests were disclosed.

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 19 March 2026

https://doi.org/10.21956/openreseurope.24541.r69741

?

Roger Huerlimann
Okinawa Institute of Science and Technology, Okinawa, Japan

This data note describes a chromosome-level reference genome assembly for *Gambusia holbrooki* (eastern mosquitofish), a globally invasive freshwater fish native to eastern North America, produced under the ERGA-BGE framework.

The male specimen was sampled from a canal in Padua, Italy, and sequenced using PacBio HiFi and Hi-C technologies. The genome was assembled using Hifiasm in Hi-C phasing mode, producing a phased, two-haplotype assembly, with the primary haplotype designated as the reference. The final assembly spans about 675 Mb organized into 24 chromosomal pseudomolecules plus a mitochondrial genome, with a contig N50 of 15.9 Mb and scaffold N50 of 29.6 Mb. BUSCO completeness at the assembly level was 99.3%, indicating a high-quality result.

Genome annotation identified 21,050 protein-coding genes and 3,687 non-coding genes, though annotation-level BUSCO completeness was lower at 88.5%. I agree with the other reviewer that this should be briefly addressed by the authors.

While this gap may partly reflect the conservative filtering steps applied during annotation, such as removal of fragmented ORFs and low-quality transcript models, the authors should briefly discuss the primary drivers of this discrepancy. Specifically, it would be useful to know whether the gap reflects limitations in the available RNA-seq data, the phylogenetic distance of the reference protein set used for gap-filling, the stringency of the Ensembl pipeline's filtering criteria, or a combination of these factors. Clarifying whether the lower annotation BUSCO score represents genuine missing gene content or is an expected outcome of conservative quality filtering would improve transparency and help readers interpret the annotation completeness metrics appropriately.

The protocols used are appropriate for this kind of work, and they are well described in the material and methods section. The data availability statement is also providing all the necessary information.

One minor comment about the first paragraph in the M&M section. I don't think it's necessary to mention ERGA's general sequencing strategy. Mentioning ONT and Illumina short read polishing even though these have not been used in this assembly might confuse readers.

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?
Yes

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

Are the datasets clearly presented in a useable and accessible format?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: My research spans marine genomics, transcriptomics, and eDNA metabarcoding in non-model fish and invertebrate species. I have hands-on experience with genome assembly and annotation, including chromosome-level assemblies of multiple fish species and the black tiger prawn, using a range of sequencing technologies and assembly approaches. I am familiar with standard quality assessment metrics including BUSCO, and with RNA-seq-based annotation workflows.

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 26 May 2026

Chiara Bortoluzzi

We thank the reviewer for their comments. We addressed as follow:

1. We thank the reviewer for their comment. As reported for reviewer #1, the annotation was generated using the automated Ensembl annotation pipeline, which applies relatively stringent filtering criteria during gene model construction. In particular, transcript-supported models are prioritised when they are additionally supported by protein homology evidence, including similarity searches against curated UniProt protein datasets. Such conservative filtering can reduce false-positive annotations but may also exclude partial, lowly expressed, lineage-specific, or weakly supported gene models, thereby lowering annotation BUSCO completeness relative to the assembly BUSCO score. In addition, the annotation relied primarily on short-read RNA-seq data, which may not fully capture transcript diversity, splice isoforms, or genes expressed only in specific tissues or developmental stages. The incorporation of broader transcriptomic resources, particularly long-read RNA sequencing and more diverse tissue sampling, could improve gene model completeness in future annotation releases. Alternative or complementary annotation strategies with less conservative filtering may also recover additional BUSCO genes
2. We removed the section on ONT data as suggested.

Competing Interests: No competing interests were disclosed.

Reviewer Report 28 February 2026

https://doi.org/10.21956/openreseurope.24541.r69744

?

Alejandro S. Mechaly
Instituto de Investigaciones en Biodiversidad y Biotecnología (INBIOTEC-CONICET), Mar del Plata, Argentina

This Data Note presents a chromosome-level reference genome assembly for *Gambusia holbrooki* generated within the ERGA-BGE framework. The manuscript is clearly written, technically rigorous, and fully consistent with the scope of a genome report. The final assembly spans 674.6 Mb organized into 24 chromosomal pseudomolecules plus the mitochondrial genome, with a contig N50 of 15.9 Mb and a scaffold N50 of 29.6 Mb. The reported BUSCO completeness of 99.3% at the assembly level and a QV of 59.8 indicate high contiguity and excellent base-level accuracy. The Hi-C contact map further supports appropriate chromosomal organization and structural reliability. Overall, the assembly meets current standards for a high-quality reference genome.

The annotation comprises 21,050 protein-coding genes and 3,687 non-coding genes. While the assembly-level BUSCO completeness is very high (99.3%), the protein-based BUSCO score for the annotation (88.5%) is noticeably lower. Although acceptable for an automated Ensembl-based pipeline, this difference is relatively pronounced for a long-read, chromosome-scale assembly of moderate genome size. The authors are encouraged to briefly discuss the reasons for this discrepancy and clarify whether it reflects annotation stringency, limited transcriptomic diversity, conservative filtering criteria, or other methodological factors. Indicating whether additional transcriptomic resources or alternative annotation strategies could improve gene completeness would strengthen the transparency of the manuscript.

The methodological description is detailed and reproducible, and sampling, vouchering, and data accessibility are well documented. The study clearly adheres to ERGA standards and provides a valuable genomic resource for a globally invasive and evolutionarily important species.

From a biological and comparative perspective, the manuscript would benefit from slightly stronger contextualization. First, incorporating reference to confirmed records of *Gambusia holbrooki* in South America, particularly Argentina—where taxonomic confusion with *G. affinis* has been documented (Cabrera et al., 2017, *Journal of Fish Biology* 91:704–710)—would strengthen the biogeographic and invasion framework of the Introduction and provide a more complete overview of the species' global introduction history.

Second, the authors may wish to position this assembly within the rapidly expanding genomic resources available for Poeciliidae. Recent chromosome-level genome assemblies for multiple *Poecilia* and *Gambusia* lineages (Ryan et al., 2025, *Genome Biology and Evolution* 17:e vaf111) highlight the growing comparative framework within the family. A concise comparison of assembly metrics and a brief discussion of how the present genome complements existing resources would help contextualize its contribution.

Overall, this is a technically robust and well-documented reference genome assembly that will serve as a useful resource for ecological, evolutionary, and comparative genomics. I recommend acceptance after minor revision.

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Yes

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

Yes

Are the datasets clearly presented in a useable and accessible format?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Specialized in fish biology, neuroendocrinology, and comparative genomics, this research focuses on the molecular and physiological mechanisms underlying reproduction and adaptation in aquatic vertebrates. Endocrine physiology is integrated with genome assembly and transcriptomic analyses to investigate sex determination, population differentiation, and responses to environmental stressors, contributing to both fundamental biological knowledge and sustainable aquaculture development.

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 26 May 2026

Chiara Bortoluzzi

We thank the reviewer for their comments. We addressed them as follow:

1. We thank the reviewer for this important observation. The discrepancy between the very high assembly-level BUSCO completeness (99.3%) and the lower protein-based BUSCO score for the annotation likely reflects characteristics of the automated annotation workflow. The annotation was generated using the automated Ensembl annotation pipeline, which applies relatively stringent filtering criteria during gene model construction. In particular, transcript-supported models are prioritised when

they are additionally supported by protein homology evidence, including similarity searches against curated UniProt protein datasets. Such conservative filtering can reduce false-positive annotations but may also exclude partial, lowly expressed, lineage-specific, or weakly supported gene models, thereby lowering annotation BUSCO completeness relative to the assembly BUSCO score. In addition, the annotation relied primarily on short-read RNA-seq data, which may not fully capture transcript diversity, splice isoforms, or genes expressed only in specific tissues or developmental stages. The incorporation of broader transcriptomic resources, particularly long-read RNA sequencing and more diverse tissue sampling, could improve gene model completeness in future annotation releases. Alternative or complementary annotation strategies with less conservative filtering may also recover additional BUSCO genes

2. We have now added the introduction of G. holbrooki in Argentina and Australia along with Europe. As well as the taxonomic confusion with G. affinis.
3. We added a reference to the GBE paper in the introduction as suggested.

Competing Interests: No competing interests were disclosed.