DATA NOTE

REVISED ERGA-BGE reference genome of *Xylophaga dorsalis* - a common deep-sea wood-boring bivalve with Atlantic-Mediterranean distribution

[version 2; peer review: 1 approved, 3 approved with reservations]

Sergi Taboada[1], Carlota Gracia-Sancha[1], Carles Galià-Camps[1], Astrid Böhne[2], Rita Monteiro[2], Thomas Marcussen[3,4], Rebekah A. Oomen[3-7], Torsten H. Struck[3], Marta Gut[8,9], Laura Aguilera[8,9], Francisco Câmara Ferreira[8,9], Fernando Cruz[8,9], Jèssica Gómez-Garrido[8,9], Tyler S. Alioto[8,9], Fergal Martin[10], Anna Lazar[10], Leanne Haggerty[10], Chiara Bortoluzzi[11]

[1]Departamento de Biodiversidad y Biología Evolutiva, CSIC, Museo Nacional de Ciencias Naturales (MNCN), Madrid, 28006, Spain
[2]Museum Koenig Bonn, Adenauerallee 127, Leibniz Institute for the Analysis of Biodiversity Change, Bonn, 53113, Germany
[3]Natural History Museum, P.O. Box 1172, Blindern, University of Oslo, Oslo, 0318, Norway
[4]Centre for Ecological & Evolutionary Synthesis, Blindernveien 31, University of Oslo, Oslo, 0371, Norway
[5]Department of Biological Sciences, 100 Tucker Park Road, University of New Brunswick Saint John, Saint John, E2K5E2, Canada
[6]Tjärnö Marine Laboratory, Hättebäcksvägen 7, University of Gothenburg, Gothenburg, 45296, Sweden
[7]Centre for Coastal Research, Universitetsveien 25, University of Agder, Kristiansand, 4630, Norway
[8]Centro Nacional de Análisis Genómico (CNAG), Baldiri Reixac 4, Barcelona, 08028, Spain
[9]Universitat de Barcelona (UB), Barcelona, 08028, Spain
[10]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, CB10 1SD, UK
[11]SIB Swiss Institute of Bioinformatics, Amphipôle, Quartier UNIL-Sorge, Lausanne, 1015, Switzerland

v2 **First published:** 30 Jan 2026, **6**:33
https://doi.org/10.12688/openreseurope.22692.1
Latest published: 11 May 2026, **6**:33
https://doi.org/10.12688/openreseurope.22692.2

Abstract

Xylophaga dorsalis is a common Atlantic-Mediterranean mollusc that plays a crucial role in deep-sea habitats, where it digests wood that reaches the seabed through a unique symbiosis with specialised bacteria. The reference genome of *X. dorsalis* thus offers a crucial resource for uncovering the genetic basis of the species adaptability to wood bore in deep-water ecosystems. The entirety of the genome sequence was assembled into 18 contiguous chromosomal pseudomolecules (superscaffolds) and 1 mitochondrial genome. This chromosome-level assembly encompasses 0.451 Gb, composed of 1,259 contigs and 320 scaffolds, with contig and scaffold N50 values of 1.30 Mb and 25.4 Mb, respectively. The genome assembly encodes 19,441 protein-coding genes (34,405 transcripts) and 6,716 non-

Open Peer Review

Approval Status ? ? ? ✓

	1	2	3	4
version 2 (revision) 11 May 2026				✓ view
version 1 30 Jan 2026	? view	? view	? view	? view

1. **Abdelmalek Lekired**, University of New Mexico, Albuquerque, USA
2. **Fabrizio Marcondes Machado**,

coding genes.

Keywords
Xylophaga dorsalis, genome assembly, European Reference Genome Atlas, Biodiversity Genomic Europe, Earth Biogenome Project, Mollusca, Bivalvia, Xylophagaidae

This article is included in the Horizon Europe gateway.

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

Universidade Estadual de Campinas, Campinas, Brazil

3. **Melody S Clark**, British Antarctic Survey, Cambridge, UK

4. **Fabrizio Ghiselli**, University of Bologna, Bologna, Italy

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: Taboada S: Investigation, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Gracia-Sancha C**: Investigation, Resources, Writing – Review & Editing; **Galià-Camps C**: 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; **Gut M**: Investigation, Writing – Review & Editing; **Aguilera L**: Investigation, Writing – Original Draft Preparation, Writing – Review & Editing; **Câmara Ferreira F**: Data Curation, Formal Analysis, Writing – Review & Editing; **Cruz F**: Data Curation, Formal Analysis, Writing – Review & Editing; **Gómez-Garrido J**: Data Curation, Formal Analysis, Writing – Review & Editing; **Alioto TS**: Data Curation, Formal Analysis, Supervision, Writing – Review & Editing; **Martin F**: Data Curation, Formal Analysis, Writing – Review & Editing; **Lazar A**: 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. ST received funding from the grant PID2020-117115GA-I00 funded by MCIN/AEI/10.13039/501100011033 and from the grant CNS2023-144572 and by the Ramón y Cajal grant RYC2021-03152-I, funded by the MCIN/AEI/10.13039/501100011033 and the European Union «NextGenerationEU»/PRTR». ST also received funding from the SponBIODIV project, funded by Biodiversa+, the European Biodiversity Partnership under the 2021 to 2022 BiodivProtect joint call for research proposals, co-funded by the European Commission (GA no. 101052342) and with the funding organizations Fundación Biodiversidad.
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: Taboada S, Gracia-Sancha C, Galià-Camps C *et al.* **ERGA-BGE reference genome of *Xylophaga dorsalis* - a common deep-sea wood-boring bivalve with Atlantic-Mediterranean distribution [version 2; peer review: 1 approved, 3 approved with reservations]** Open Research Europe 2026, **6**:33 https://doi.org/10.12688/openreseurope.22692.2

First published: 30 Jan 2026, **6**:33 https://doi.org/10.12688/openreseurope.22692.1

REVISED **Amendments from Version 1**

In this revised version, we have addressed all reviewers comments, which have resulted in an extension of the genome assembly methods and results section. We have also corrected all inconsistencies reported in the initial version, which were mostly linked to the genome annotation section. The introduction also presents some new rewording and contextualisation of the here produced genome assembly.

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

Introduction

Xylophaga dorsalis, a mollusc from the family *Xylophagaidae*, is a deep-sea wood-boring bivalve commonly found in association with wood remains at depths ranging from 130 to 2,000 m depth (Romano *et al*., 2014). *Xylophaga dorsalis* plays a crucial role in the deep sea, boring into and digesting wood that reaches the seabed, thus accelerating its decomposition. This activity releases recalcitrant organic matter retained in the wood, putting back nutrients into the benthic ecosystems, supporting diverse communities of opportunistic microbes and scavengers. Importantly, the burrowing of these bivalves creates complex ephemeral microhabitats that increase local biodiversity (Bienhold *et al*., 2013). In this respect, *Xylophaga* species are functionally analogous to shallow-water wood-boring bivalves such as teredinids, although they operate in deep-sea environments where wood falls represent a major, yet patchy, source of organic input. While *X. dorsalis* is distributed across the Mediterranean and in the North-East Atlantic from the Cantabrian Sea to the northern part of Norway, previous reports of its occurrence in the North-West Atlantic are now in doubt (Romano *et al*., 2014).

During evolution *Xylophaga* species have established a unique relationship with endosymbiotic bacteria found in the gills, which convert otherwise unavailable energy sources such as cellulose into forms readily metabolized by the host (Distel & Roberts, 1997). Developing a high-quality reference genome for *X. dorsalis* is thus essential for advancing our understanding of its unique genetic makeup and adaptive traits and will shed light onto host-symbiont coevolutionary processes. A reference genome will thus provide the means to identify the genetic basis of metabolic complementarity between host and symbiont, offering insights into how these partnerships have evolved and are maintained.

The generation of this reference resource, which is the first and only one available for the *Xylophaga* genus as retrieved from Genomes on a Tree (Challis *et al*., 2023), 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 Oxford Nanopore Technology (ONT) and/or Pacific Biosciences (PacBio) for long-read sequencing, along with Hi-C sequencing for chromosomal architecture, Illumina Paired-End (PE) for polishing (i.e. recommended for ONT-only assemblies), and RNA sequencing for transcriptomic profiling, to facilitate genome assembly and annotation.

Sample and sampling information

On 28th of May 2023, Sergi Taboada and Carlota Gracia-Sancha sampled 50 specimens of *Xylophaga dorsalis* (sex unknown). Specimens were collected from a sunken tree log at a depth of 1,231 m near the shore of Mazarrón, Murcia, Spain, during the *LanderFleet Mazarrón 2023* sampling campaign aboard the Ramón Margalef vessel. Specimens were identified by Sergi Taboada based on barcoding a fragment of the cytochrome *c* oxidase subunit I gene and morphology following Romano *et al*. (2014). The biological material collected in Spain, and used to generate digital sequences, was retrieved from wildlife taxa regulated by the Spanish Royal Decree 124/2017. No ABS permits were granted as utilization of the material falls under the exemption in Article 3(2) of the Spanish Royal Decree 124/2017 and was confirmed by the national authorities. Specimens were caught with forceps and were euthanized by snap-frozing them directly in liquid nitrogen. Specimens were kept in liquid nitrogen until DNA extraction.

Vouchering information

Physical reference materials for the sequenced specimen have been deposited in the Museo Nacional de Ciencias Naturales (MNCN-CSIC) https://www.mncn.csic.es/es under the accession number 15.07/20008.

Reference tissue material of the body wall is available from a proxy individual at the Biobank of the Museo Nacional de Ciencias Naturales (MNCN-CSIC) https://www.mncn.csic.es/es under the voucher ID MNCN-ADN 151.772.

Genetic information

According to data retrieved from Genomes on a Tree (Challis *et al.*, 2023), the genome was expected to be diploid, with an expected haploid number of 19 chromosomes (2n = 38) and an estimated genome size based on the ancestral taxa *Cyrtopleura costata* of 0.68 Gb. This estimated genome size was used prior to sequencing to help calculate sequencing effort. After sequencing, counting of k-mers in the Illumina read data with Meryl (k = 20) (Rhie *et al.*, 2020) and genome size estimation with Genomescope 2.0 (Ranallo-Benavidez *et al.*, 2020) gives an estimated haploid genome size of 454 Mb.

DNA/RNA processing

DNA was extracted from one whole specimen using the Blood & Cell Culture DNA Mini Kit (Qiagen) following the manufacturer's instructions. DNA quantification was performed using a Qubit dsDNA BR Assay Kit (Thermo Fisher Scientific), and DNA integrity was assessed using a Genomic DNA 165 Kb Kit (Agilent) on the Femto Pulse system (Agilent). The DNA was stored at +4 °C until sequenced.

RNA was extracted from the whole organism using a RNeasy Mini Kit (Qiagen) according to the manufacturer's instructions. RNA quantification was performed using the Qubit RNA BR kit, and RNA integrity was assessed using a Bioanalyzer 2100 system (Agilent) RNA 6000 Nano Kit (Agilent). The RNA was stored at −80 °C until sequenced.

Library preparation and sequencing

For long-read whole genome sequencing, a library was prepared using the SQK-LSK114 Kit (Oxford Nanopore Technologies, ONT), which was then sequenced in one R10.4.1 flow cell on a PromethION 24 A Series instrument (ONT). A short-read whole-genome sequencing library was prepared using the KAPA Hyper Prep Kit (Roche). A Hi-C library was prepared from the whole specimen using the Dovetail Omni-C Kit (Cantata Bio), followed by the KAPA Hyper Prep Kit for Illumina sequencing (Roche). The RNA library was prepared using the KAPA mRNA Hyper prep kit (Roche). All short-read libraries were sequenced on a NovaSeq 6000 instrument (2x150bp, Illumina). In total 93x Oxford Nanopore, 230x Illumina WGS shotgun, and 178x HiC data were sequenced to generate the assembly.

Genome assembly methods

The genome was assembled using the CLAWS pipeline (Gomez-Garrido, 2024). Briefly, reads were pre-processed for quality and length using Trim Galore v0.6.7 (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) and Filtlong v0.2.1 (https://github.com/rrwick/Filtlong), and initial contigs were assembled using Flye v2.9.1-b1780 (Kolmogorov *et al.*, 2019) with default parameters, followed by polishing of the assembled contigs using HyPo v1.0.3 (Kundu *et al.*, 2019), and removal of retained haplotigs using purge-dups v1.2.6 (Guan *et al.*, 2020). Six short contigs (<100 kb) corresponding to bacterial sequences (Pseudomonadota, Bacillota) were identified with BlobToolKit v0.6.0 (Challis *et al.*, 2020) and removed from the assembly. The scaffolded assembly, which spans a total length of 451,816,200 bp, was obtained by scaffolding the contig-level assembly with Hi-C data using YAHS v1.2a (Zhou *et al.*, 2023) followed by manual curation in PretextView v0.2.5 (https://gitlab.com/wtsi-grit/rapid-curation) to remove any false joins and incorporate any sequences not automatically scaffolded into their respective locations in the chromosomal pseudomolecules (or super-scaffolds). Finally, the mitochondrial genome was assembled using ONT reads as a single circular contig of 21,574 bp using the FOAM pipeline (https://github.com/cnag-aat/FOAM) and included in the released assembly (GCA_965225065.2). BUSCO v5.8.0 was used to estimate the level of genome completeness on the final assembly using the Eukaryota database (odb10), whereas Merqury v1.3 (Rhie *et al.*, 2020) was used to estimate base accuracy Quality Value. Summary analysis of the released assembly was performed using the ERGA-BGE Genome Report ASM Galaxy workflow (https://doi.org/10.48546/workflowhub.workflow.1103.2).

Genome annotation methods

A gene set was generated for a previously released chromosome-level genome assembly (GCA_965225065.1) using the Ensembl Gene Annotation system (Aken *et al.*, 2016), primarily by aligning publicly available short-read RNA-seq data from BioSample SAMEA116055609 to the genome. Gaps in the annotation were filled via protein-to-genome alignments of a select set of clade-specific 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 metazoan 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 two categories: protein-coding, and long non-coding. Models that 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 was assembled with Flye, polished, purged and scaffolded and curated as detailed in the methods. The curated assembly comprises 18 superscaffolds and 286 scaffolds that could not be placed into any superscaffold, and the mitogenome (Figures 1 & 2). The assembly has a GC content of 38.1%, a contig N50 of 1,253,263 bp (L50 = 85), and a scaffold N50 of 25,438,142 bp (L50 = 8). A total of 939 gaps are present, with a cumulative gap size of 187.8 kb. Completeness assessed using BUSCO (Manni *et al.*, 2021) against the Eukaryota database (odb10) was 97.6% (95.3%

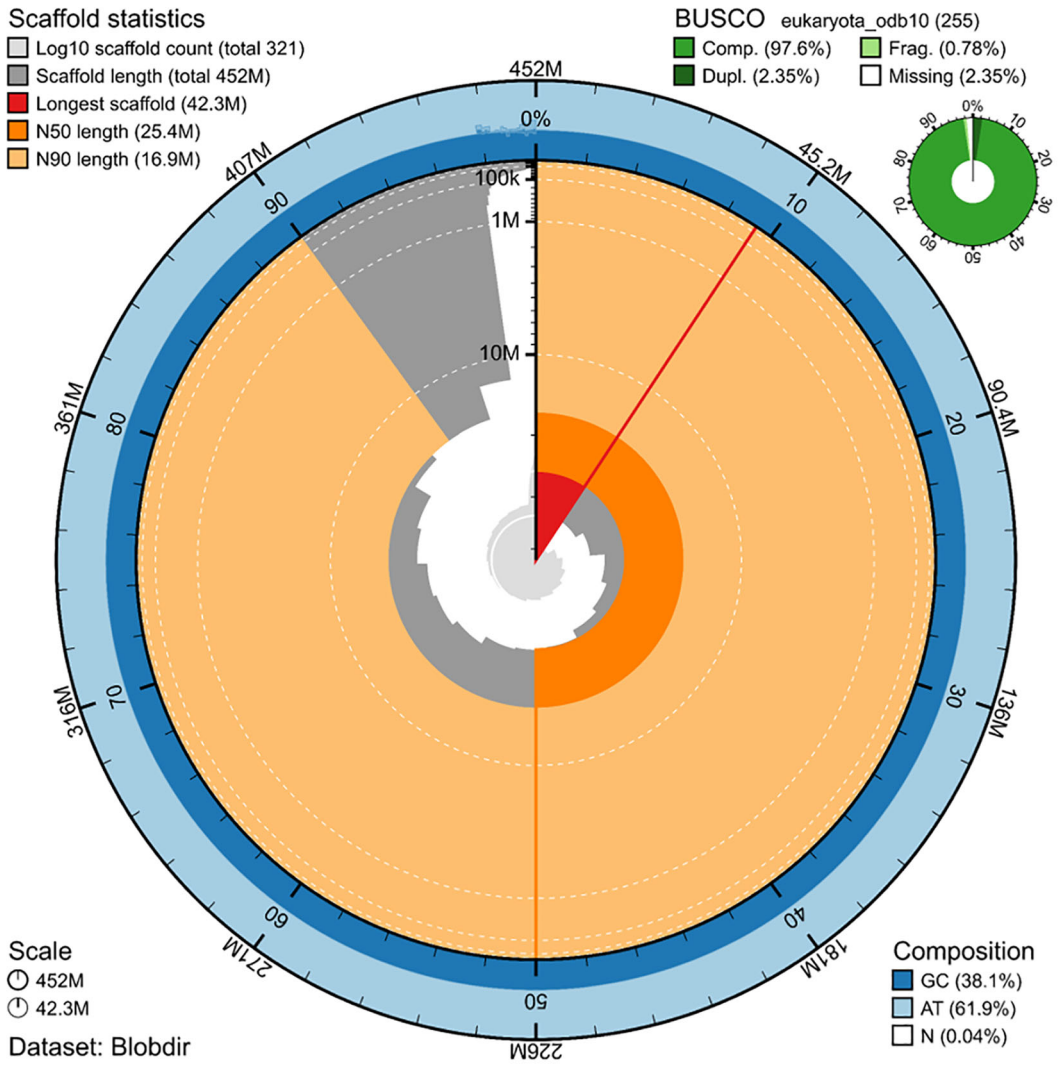

Figure 1. Snail plot summary of assembly statistics. The main plot is divided into 1,000 size-ordered bins around the circumference, with each bin representing 0.1% of the 451,816,200 bp assembly including the mitochondrial genome. The distribution of sequence lengths is shown in dark grey, with the plot radius scaled to the longest sequence present in the assembly (42.3 Mb, shown in red). Orange and pale-orange arcs show the scaffold N50 and N90 sequence lengths (25,438,142 and 16,851,157 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 Eukaryota database (odb10) is shown in the top right.

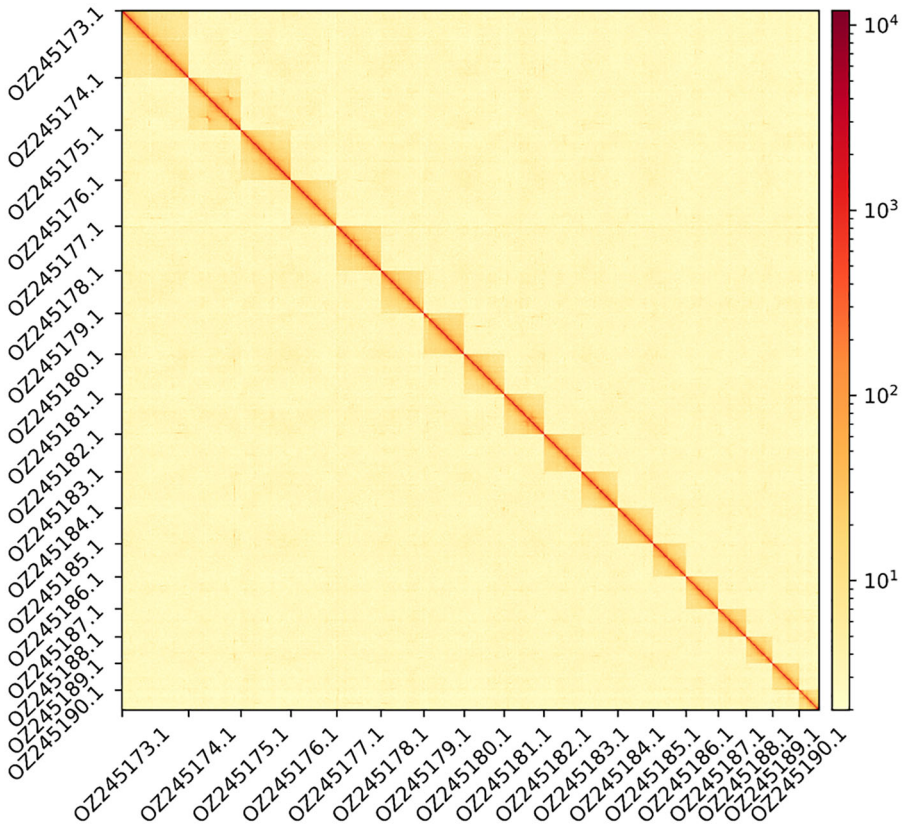

Figure 2. Hi-C contact map showing spatial interactions between regions of the genome. The diagonal corresponds to intra-chromosomal contacts, depicting chromosome boundaries. The frequency of contacts is shown on a logarithmic heatmap scale. Hi-C matrix bins were merged into a 300 kb bin size for plotting. For visualisation purposes, the Hi-C contact map shows only the 18 chromosomes.

single-copy and 2.4% duplicated). Finally, the assembly has a base accuracy Quality Value (QV) or 49.7 and a kmer completeness of 68% as calculated by Merqury (Rhie *et al.*, 2020).

Genome annotation

The genome annotation consists of 19,441 protein-coding genes with associated 34,405 transcripts, in addition to 6,716 non-coding genes (Table 1). BUSCO was run using the Mollusca database (odb12) on the canonical proteins to avoid high duplication caused by isoforms. This led to a completeness score of 95.9%. Using the OMAmer Metazoa-v2.0.0.h5 database for OMArk (Nevers *et al.*, 2025) resulted in 94.1% completeness and 58.2% consistency (Table 2).

Table 1. Statistics from assembled gene models.

	No. genes	No. transcripts	Mean gene length (bp)	No. single-exon genes	Mean exons per transcript
mRNA	19,441	34,405	15,708	568	9.9
pseudogene	0.0	0.0	0.0	0.0	0.0
snoRNA	162	162	112	162	1.0
lncRNA	5,237	6,113	5,253	54	2.3
tRNA	1,046	1,046	75	1,046	1.0
snRNA	128	128	152	128	1.0
rRNA	141	141	407	141	1.0
scRNA	2	2	132	2	1.0

Table 2. Annotation completeness and consistency scores calculated by BUSCO run in protein mode (mollusca_odb12) and OMArk (Metazoa-v2.0.0.h5).

	Complete	Singular	Duplicated	Fragmented	Missing
BUSCO	4,241 (95.9%)	4,173 (94.4%)	68 (1.5%)	116 (2.6%)	64 (1.4%)
OMArk	2,027 (94.1%)	364 (16.9%)	1,663 (77.2%)	-	126 (5.8%)
	Consistent	**Inconsistent**	**Contaminants**	**Unknown**	
OMArk	23,118 (58.2%)	4,160 (10.5%)	0 (0.0%)	12,442 (31.3%)	

Data availability

Xylophaga dorsalis and the related genomic study were assigned to Tree of Life ID (ToLID) 'xbXylDors1' and all sample, sequence, and assembly information are available under the umbrella BioProject PRJEB87517. The sample information is available at the following BioSample accessions: SAMEA116055604, SAMEA116055609, and SAMEA116055611. The genome assembly is accessible from ENA under accession number GCA_965225065.2. This version includes the mitogenome, while the first version GCA_965225065.1 includes only the nuclear genome. The gene annotation is on the Ensembl webpage (https://projects.ensembl.org/erga-bge/). Sequencing data produced as part of this project are available from ENA at the following accessions: ERX14171855, ERX14171856, ERX14171857, and ERX14171858. 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/blob/main/Assembly_Reports/Xylophaga_dorsalis/xbXylDors1/. Further details and data about the project are hosted on the ERGA portal at https://portal.erga-biodiversity.eu/data_portal/1526741.

Author contributions

ST coordinated the project; ST, CG-S and CG-C collected the species; ST, CG-S and CG-C identified the species; ST, CG-S and CG-C 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; LA and MG extracted DNA, prepared libraries, and performed sequencing; FCF, FC, JG-G, and TSA performed genome assembly and curation under the supervision of TSA; FM, AL, 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 would like to express our gratitude to the crew of the R/V Ramón Margalef for their help during sampling. Special thanks to Eneko Aierbe and César González-Pola as the researchers of the Instituto Español de Oceanografía (IEO-CSIC) in charge of the LanderFleet project. We also thank all members of the Riesgo and Taboada's labs for their assistance during sampling preparation. 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, Jo Collins from the Wellcome Sanger Institute.

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

Bienhold C, Pop Ristova P, Wenzhöfer F, *et al.*: **How deep-sea wood falls sustain chemosynthetic life.** *PLoS One.* 2013; **8**(1): e53590.
PubMed Abstract | Publisher Full Text | Free 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

Challis R, Richards E, Rajan J, *et al.*: **BlobToolKit-interactive quality assessment of genome assemblies.** *G3 (Bethesda).* 2020; **10**(4): 1361–1374.
PubMed Abstract | Publisher Full Text | Free Full Text

Distel DL, Roberts SJ: **Bacterial endosymbionts in the gills of the deep-sea wood-boring bivalves *Xylophaga atlantica* and *Xylophaga washingtona*.** *Biol Bull.* 1997; **192**(2): 253–261.
PubMed Abstract | Publisher Full Text

Gomez-Garrido J: **CLAWS (CNAG's long-read assembly workflow in Snakemake).** [computer software]. 2024.
Publisher Full Text

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

Guan D, McCarthy SA, Wood J, *et al.*: **Identifying and removing haplotypic duplication in primary genome assemblies.** *Bioinformatics.* 2020; **36**(9): 2896–2898.
PubMed Abstract | Publisher Full Text | Free Full Text

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**

Kolmogorov M, Yuan J, Lin Y, *et al.*: **Assembly of long, error-prone reads using repeat graphs.** *Nat Biotechnol.* 2019; **37**(5): 540–546.
PubMed Abstract | **Publisher 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**

Kundu R, Casey J, Sung WK: **HyPo: super fast & accurate polisher for long read genome assemblies.** *bioRxiv.* 2019; **2019**(12)
Publisher Full Text

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.
PubMed Abstract | **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**

Ranallo-Benavidez TR, Jaron KS, Schatz MC: **GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes.** *Nat Commun.* 2020; **11**(1): 1432.
PubMed Abstract | **Publisher Full Text** | **Free 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**

Romano C, Voight JR, Perez-Portela R, *et al.*: **Morphological and genetic diversity of the wood-boring Xylophaga (Mollusca, Bivalvia): new species and records from deep-sea Iberian canyons.** *PLoS One.* 2014; **9**(7): e102887.
PubMed Abstract | **Publisher Full Text** | **Free 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**

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 13 May 2026

https://doi.org/10.21956/openreseurope.25825.r74036

✓ **Fabrizio Ghiselli**
University of Bologna, Bologna, Italy

The Authors have addressed the concerns I raised in my peer review report appropriately.

I have no further comments to make.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Comparative Genomics, Genome Evolution, Molecular Evolution, Mitochondria, Biodiversity Genomics

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.24540.r70658

? **Fabrizio Ghiselli**
University of Bologna, Bologna, Italy

This manuscript presents a chromosome-level genome assembly of *Xylophaga dorsalis*, a species

of ecological importance due to its role in wood degradation in deep-sea ecosystems. The dataset represents a valuable resource for the research community. Several aspects of the manuscript would benefit from clarification and additional details. Addressing these points would substantially strengthen the work and improve its clarity and reproducibility prior to publication.

I share several of the concerns outlined by Reviewers 1 and 3. My comments are provided below.

- Genome Size Estimation and Assembly Validation

The genome size of *X. dorsalis* is estimated based on a closely related species. However, given the increasing evidence of high variability in genome size even among closely related species (and even within species), extrapolation from related taxa is not recommended. Genome size and ploidy should be estimated directly from the sequencing data using k-mer–based approaches (e.g., GenomeScope or similar).

In addition, it would be useful to provide a table including the following:

A k-mer profile.
Coverage statistics.
Heterozygosity estimates.
Assembly quality metrics (e.g., N50, L50, total assembly size, number of contigs/scaffolds).
BUSCO results, standardized using a single database version and clearly linked to the assembly version analyzed.

If Merqury was used, full statistics should be described in the Methods.

Most of these metrics appear to be shown in the snail plot; however, a summary table would be easier to read and interpret.

- Inconsistencies in Annotation Statistics: as noted by Reviewers 1 and 3, I also observed inconsistencies between the annotation statistics reported in the Abstract and those presented in the Results and Table 1. These discrepancies should be carefully reviewed and corrected to ensure consistency throughout the manuscript.

- Transcriptome Results Should Be Described in Greater Detail. Additional information would greatly improve clarity and reproducibility.

For example:

The number of RNA-seq samples used.
Assembly statistics.
Whether genome-guided and/or de novo transcriptome assemblies were performed.
How RNA-seq data were integrated into gene model construction.
Biosample IDs (e.g., SAMEA116055609), which could not be accessed or verified.
Total reads per sample.
Assembly size and N50 (if de novo assembled).
BUSCO completeness of the transcriptome.

Mapping rates to the genome.
Providing these details would strengthen the annotation section considerably.

- Mitochondrial Genome

Regarding Reviewer #1's concern about the mitochondrial genome size appearing larger than expected compared to published *X. dorsalis* mitochondrial data (~16.8 kb), I would like to add a comment.

I have worked on bivalve mitochondrial genomics for 20 years, and mitochondrial genome size can be quite variable, even within species, often due to repeats localized upstream or downstream of the control region. It is also important to consider that this study used long-read sequencing. If the previously published genome was generated using short-read data, repeat regions may have been collapsed, resulting in an underestimated genome size. Therefore, the observed size difference may reflect biological reality rather than an assembly artifact.
That said, given the peculiarities of mitochondrial biology in bivalves, Reviewer #1's suggestions for careful verification are appropriate. There is also an additional related issue to consider. Many (certainly over 100) bivalve species exhibit Doubly Uniparental Inheritance (DUI) of mitochondria. DUI species possess two distinct mitochondrial lineages with separate inheritance patterns (maternal and paternal). These lineages can be highly divergent (up to 0.45 nucleotide p-distance), raising the possibility of chimeric assemblies or other assembly complications. DUI can be difficult to detect. In some cases, it is suggested by unusually high heterozygosity in mitochondrial sequences; in others, two clearly distinct mitochondrial genomes may assemble separately.

While assessing DUI is beyond the scope of this genome report, if the authors identify any signals or indications suggesting its presence, reporting this would be valuable for the community.

I also found that the links to the MNCN website do not function, and voucher specimens could not be located using the search tool.

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?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Comparative Genomics, Genome Evolution, Molecular Evolution, Mitochondria, Biodiversity Genomics

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 30 Apr 2026

Chiara Bortoluzzi

We thank the reviewer for their comments. We have addressed all of them as below:

1. We understand the confusion. The genome size estimated using related species was done only to estimate the amount of sequencing required prior to sequencing. We rephrase this in the main text.
2. All of these statistics are already reported in the ERGA Assembly Report (EAR, https://github.com/ERGA-consortium/EARs/blob/main/Assembly_Reports/Xylophaga_dorsalis/xbXylDors1/) cited in the manuscript. Hence, we did not see the point of reporting this information once again as the reader is referred to the EAR in the main text.
3. We now report Merqury in our Materials and methods section
4. We have now corrected the genome annotation results throughout the manuscript
5. The text reported in the main text for the genome annotation section follows the standard Ensembl template. It follows that most Ensembl results are reported in their website. In the case of *X. dorsalis*, the full annotation statistics can be found here: https://beta.ensembl.org/species/84d8aa9e-1f9e-40fe-a304-b3d2d5df1e9a. Regarding BioSample SAMEA116055609, this can be accessed here: https://www.ebi.ac.uk/biosamples/samples/SAMEA116055609.
6. The links to the MNCN are meant to redirect the reader to the National History Museum and not to the voucher itself. To access the voucher a request should be submitted to the Museum.

Competing Interests: No competing interests were disclosed.

Reviewer Report 12 March 2026

https://doi.org/10.21956/openreseurope.24540.r70657

?

Melody S Clark
British Antarctic Survey, Cambridge, UK

The ecological role of Xylophaga dorsalis is well described and the genome has been subjected to comprehensive sequencing. However, I have a number of concerns.

Is this an updated genome assembly based on GCA_965225065.1? This is unclear, whether this is an updated reference genome based on these previous data or a completely separate one?

Firstly none of the links to the mncn website work in terms of access to the information on the voucher specimens. I could not find the specimens using the mncn search function. There needs to be pictures of the voucher specimens included in this publication and links to the specimens in mncn need to be direct links.

The genome size of this species is estimated using related species, but this must be based on actual data (k-mer-based analysis as per Reviewer 1), as I am sure the authors are aware of the c-value paradox.

The data from each individual sequencing method must be presented in a table.

The workflows are not clear. Links to centralised generic workflows with multiple pathways are very confusing and the authors need to explicitly state how they generated both the genome assembly and the annotation, so that these can be replicated.

More data are needed on the RNA-seq: assembly stats etc, especially as I could not access the Biosample data typing in the IDs provided in this publication e.g SAMEA116055609. Furthermore, I am not convinced about the use of clade-specific proteins for annotation. The authors should generate more RNA-seq data if required for annotation.

The data in Table 1 do not correlate with the abstract and needs correction.

The authors state that they generated 320 scaffolds, but these ultimately resulted in 18 chromosomal psuedochromosomes, plus the mitochondrial genome. The authors need to state the the contig and scaffold data for each pseudochromosome.

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?

No

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

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Transcriptomics and genomics of non-model species

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 30 Apr 2026

Chiara Bortoluzzi

We thank the reviewer for their comments. We have addressed all of them as below:

1. The genome assembly for which we report on is an improved version of GCA_965225065.1. We specified this in the main text.
2. The links to the MNCN are meant to redirect the reader to the National History Museum and not to the voucher itself. To access the voucher a request should be submitted to the Museum.
3. We understand the confusion on the estimated genome size. The genome size estimated via GoaT was used to estimate the amount of sequencing required prior to sequencing.
4. As suggested, we expanded the genome assembly section to ensure replicability. With regard to the genome annotation part, the text reported in the main text follows the standard Ensembl template.
5. All BioSamples data can be access either via ENA (https://www.ebi.ac.uk/ena/browser/view/SAMEA116055609) or BioSamples (https://www.ebi.ac.uk/biosamples/samples/SAMEA116055609). We cannot generate additional RNA-Seq data because the BGE project has ended as well as funding.
6. We indeed realised the mistake and have now corrected the Abstract to meet the results presented in Table 1.
7. Following Hi-C mapping to a flye assembly and manual curation in Pretext View, the final scaffolded assembly submitted to the ENA comprises 18 SUPER scaffolds (pseudochromosomes) and 286 smaller scaffolds that could not be placed onto any pseudochromosome, for a total of 320 sequences. We specified this in the main text.

Competing Interests: No competing interests were disclosed.

Reviewer Report 05 March 2026

https://doi.org/10.21956/openreseurope.24540.r70656

?

Fabrizio Marcondes Machado
Universidade Estadual de Campinas, Campinas, Brazil

The study represents a valuable genomic resource for marine biology and deep-sea ecology. The chromosome-level assembly is technically robust, with consistent N50 statistics, and the gene annotation provides a high-density dataset for future functional and comparative studies. The

work addresses an ecologically significant species and offers insights into the genetic basis of its symbiosis and wood-digestion capabilities.

Discussion: However, some aspects require clarification or further elaboration. Some numbers, while plausible, could benefit from additional context or comparison with related bivalves. Although this is a study with a very specific focus, it is not clear whether a complete genome of another species in the same genus or family already exists. Furthermore, while the manuscript mentions the ecological role of *X. dorsalis*, it could benefit from a comparative perspective with other deep-sea bivalves or wood-boring species, perhaps even regarding the number of genes identified.

Introduction: Although I understand how a high-quality genome can help us understand host-symbiont coevolutionary processes, I would suggest that the authors include a brief explanatory sentence here. I am confident that doing so would benefit readers, especially younger researchers, by clarifying why this genome is particularly valuable for studying these interactions.

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: Morphology and anatomy of bivalves, Taxonomy of marine molluscs, Genomics of bivalves.

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 30 Apr 2026

Chiara Bortoluzzi

We thank the reviewer for their comments. We have addressed all of them as below:

1. We now specify in the introduction that the reference genome of *X. dorsalis* is the first and only one available for the *Xylophaga* genus.
2. We now included a brief explanatory sentence in the Introduction explaining how a high-quality genome can help us understand host-symbiont co-evolutionary processes.

Competing Interests: No competing interests were disclosed.

Reviewer Report 10 February 2026

https://doi.org/10.21956/openreseurope.24540.r69058

?

Abdelmalek Lekired
University of New Mexico, Albuquerque, NM, USA

Reviewer Report:

We thank the authors for this valuable contribution to molluscan genomics, particularly for marine and deep-sea molluscs, which remain underrepresented in genomic datasets. The authors present a chromosome-level assembly and annotation of *Xylophaga dorsalis* as a data note. Overall, the work is potentially valuable and merits publication; however, several concerns regarding the methodology and the consistency of the reported data must be addressed before acceptance.
Major Comments

Abstract

- The authors report 18 chromosomes and one mitochondrial genome; however, key genomic metrics are missing, including genome size, sequencing coverage, heterozygosity, BUSCO completeness, and general assembly statistics.
- Transcriptome metrics are not provided, and the annotation numbers reported in the abstract are inconsistent with those presented in the Results section.
- It is unclear whether the annotation was performed on the chromosome-level assembly or on a scaffold-level assembly.
- The number of contigs is mentioned in the abstract but not clearly reported in the Results section.

Recommendation: All reported metrics should be based consistently on the chromosome-level assembly and clearly stated.

Results – Genome Assembly

- The estimated genome size based on a closely related species is not sufficiently reliable. Genome size and ploidy should be estimated directly using k-mer–based analyses.
- Assembly statistics (e.g., N50, total assembly size, coverage, BUSCO, Merqury if used) should be clearly reported in Material and methods.

Genome Annotation

- There is a major inconsistency between annotation statistics reported in the abstract and

those in the Results:
    - **Abstract:** 39,720 protein-coding genes (70,084 transcripts) and 12,153 non-coding genes.
    - **Results:** 19,441 protein-coding genes with 34,405 transcripts and 6,716 non-coding genes.

These discrepancies must be resolved and explained.

- The annotation methodology is not described clearly enough to be reproducible. Critical details are missing, including:
    - Whether the genome was soft-masked prior to annotation.
    - Which genome assembly version (contig, scaffold, or chromosome level) was used.
    - Whether ab initio gene predictors (e.g., AUGUSTUS) or RNA-Seq–based annotation tools were used.
- RNA-Seq usage is unclear:
    - How many RNA-Seq samples were used?
    - Were transcript models generated using genome-guided and/or de novo transcriptome assemblies?
    - How were these data integrated into the annotation?
- The Methods section reports results that are not described methodologically, including BUSCO and Merqury analyses.
- Two different BUSCO database versions are mentioned (odb10 for Metazoa and odb12 for Mollusca) without justification. This needs clarification and standardization.

Recommendation: Provide a clear, step-by-step annotation pipeline, ideally supported by a flowchart, and list all software tools and versions used.

Mitochondrial Genome

- The reported mitochondrial genome size appears suspiciously large, exceeding the known *Xylophaga dorsalis* mitochondrial genome by ~5 kb (GenBank: OM910830.1; 16,787 bp).
- The mitochondrial assembly should be carefully checked for duplication or assembly artifacts.

Contamination Assessment

- The authors report only six short bacterial contigs identified using BlobToolKit, but it is unclear:
    - Which assembly version was used for this analysis (contigs, scaffolds, or chromosome-level assembly)?
    - How contaminants were identified and removed in practice.
- It should be clarified whether decontamination was performed before or after chromosome-level scaffolding.

Missing or Unclear Information

The manuscript would benefit from the following clarifications and additions:

1. A chronological and clearly described genome assembly workflow.
2. A reproducible genome annotation pipeline, including a flowchart.
3. Summary annotation statistics per chromosome.
4. The draft genome size prior to BlobToolKit filtering and chromosome-level assembly.
5. The number/percentage of contigs/sequences not anchored to chromosomes.

Overall Recommendation

The dataset has clear potential value for the community; however, substantial revisions are required to improve methodological transparency, data consistency, and reproducibility. Addressing the points above will significantly strengthen the manuscript and its utility as a genomic resource.

References

1. Lekired A, Mudge J, Laidemitt M, Mutuku M, et al.: Whole-genome sequencing and genome-wide transcriptome profiling of the freshwater planorbid snail Bulinus ugandae (Mollusca: Gastropoda), a Nilotic bulinine refractory to Schistosoma haematobium. *BMC Genomics*. 2025; **26** (1). Publisher Full Text
2. Lekired A, Cherif-Silini H, Silini A, Ben Yahia H, et al.: Comparative genomics reveals the acquisition of mobile genetic elements by the plant growth-promoting Pantoea eucrina OB49 in polluted environments. *Genomics*. 2023; **115** (2). Publisher Full Text
3. Habib M, Posavi M, Lekired A, Zhang S: Exploring the genome-wide transcriptomic responses of Bulinus truncatus to Schistosoma haematobium infection: An important host-parasite system involved in the transmission of human urogenital schistosomiasis. *Molecular Immunology*. 2024; **175**: 74-88 Publisher Full Text
4. Benyedem H, Lekired A, Mhadhbi M, Dhibi M, et al.: First insights into the microbiome of Tunisian Hyalomma ticks gained through next-generation sequencing with a special focus on H. scupense. *PLOS ONE*. 2022; **17** (5). Publisher Full Text
5. Zhang S, Yan G, Lekired A, Zhong D: Genomic basis of schistosome resistance in a molluscan vector of human schistosomiasis. *iScience*. 2025; **28** (1). Publisher Full Text

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

Partly

Are the protocols appropriate and is the work technically sound?

Yes

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

No

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

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: My expertise is in genome assembly, curation, and functional annotation of bacterial, fungal, and molluscan genomes, with a particular focus on non-model organisms. I have extensive experience generating and evaluating hybrid genome assemblies using short-read and long-read sequencing data, and integrating RNA-seq evidence to support gene prediction and transcript validation. My work includes assembly quality assessment, resolution of complex genomic features such as repeats, heterozygosity, and ploidy, and evidence-based gene annotation using RNA-seq–guided pipelines. I routinely analyze RNA-seq data for genome annotation, expression profiling, and locus-level interpretation, with an emphasis on producing biologically robust genome resources suitable for comparative and functional genomics.

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.