Check for updates

DATA NOTE

REVISED ERGA-BGE reference genome of the lineid heteronemertean *Lineus lacteus* (Pilidiophora, Nemertea)

[version 2; peer review: 2 approved]

Aida Verdes[1], Patricia Alvarez-Campos[2], María Conejero[1], Ana Riesgo[1], Astrid Böhne[3], Rita Monteiro[3], Javier Palma-Guerrero[4,5], Rosa Fernández[5], Marta Gut[6,7], Laura Aguilera[6,7], Francisco Câmara Ferreira[6,7], Fernando Cruz[6,7], Jèssica Gómez-Garrido[6,7], Tyler S. Alioto[6,7], Chiara Bortoluzzi[8]

[1]Department of Biodiversity and Evolutionary Biology, Calle José Gutiérrez Abascal 2, National Museum of Natural Sciences (MNCN-CSIC), Madrid, 28006, Spain
[2]Department of Biology, Calle Darwin 2, Autonomous University of Madrid, Madrid, 28049, Spain
[3]Leibniz Institute for the Analysis of Biodiversity Change, Adenauerallee 127, Museum Koenig Bonn, Bonn, 53113, Germany
[4]Research Institute of Organic Agriculture, Ackerstrasse 113, FiBL, Frick, 5070, Switzerland
[5]Metazoa Phylogenomics Lab, Passeig marítim de la Barceloneta 37-49, Institute for Evolutionary Biology (CSIC-UPF), Barcelona, 08003, Spain
[6]Centro Nacional de Análisis Genómico (CNAG), Baldiri Reixac 4, Barcelona, 08028, Spain
[7]Universitat de Barcelona (UB), Barcelona, 08028, Spain
[8]SIB Swiss Institute of Bioinformatics, Amphipôle, Quartier UNIL-Sorge, Lausanne, 1015, Switzerland

v2 **First published:** 16 Oct 2025, **5**:317
https://doi.org/10.12688/openreseurope.21501.1
Latest published: 25 Nov 2025, **5**:317
https://doi.org/10.12688/openreseurope.21501.2

Abstract

The reference genome of *Lineus lacteus* is a crucial resource for studying the genetic basis of novelty and the evolution of remarkable traits, such as regeneration and venom, as well as the molecular mechanisms underlying adaptability in marine intertidal ecosystems. *Lineus lacteus* belongs to the Nemertea, a phylum of worm-shaped animals comprising approximately 1,300 species within the Lophotrochozoa — a superphylum of animals including leeches, snails, and other invertebrates that is crucial to our understanding of bilaterian evolution. Despite their evolutionary and ecological relevance, genomic resources for the phylum Nemertea remain scarce. We assembled the entirety of the *L. lacteus* genome into 19 contiguous chromosomal pseudomolecules. This chromosome-level assembly encompasses 0.37 Gb, composed of 71 contigs and 27 scaffolds, with contig and scaffold N50 values of 8.9 Mb and 20.4 Mb, respectively.

Open Peer Review

Approval Status ✓ ✓

	1	2
version 2 (revision) 25 Nov 2025		✓ view
version 1 16 Oct 2025	✓ view	

1. **Thomas D Lewin**, Biodiversity Research Center Academia Sinica, Taipei City, Taiwan

2. **Natsumi Hookabe**, Japan Agency for Marine-Earth Science and Technology (JAMSTEC), Yokosuka, Japan

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

Keywords
Lineus lacteus, Lineidae, genome assembly, European Reference Genome Atlas, Biodiversity Genomics Europe, Earth Biogenome Project, nemertean, ribbon worm

This article is included in the Horizon Europe gateway.

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

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

Author roles: Verdes A: Investigation, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Alvarez-Campos P**: Investigation, Resources, Writing – Review & Editing; **Conejero M**: Investigation, Resources, Writing – Review & Editing; **Riesgo A**: Investigation, Resources, Writing – Review & Editing; **Böhne A**: Methodology, Project Administration, Supervision, Writing – Review & Editing; **Monteiro R**: Methodology, Project Administration, Supervision, Writing – Review & Editing; **Palma-Guerrero J**: Methodology, Project Administration, Supervision, Writing – Review & Editing; **Fernández R**: 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; **S. Alioto T**: Data Curation, Formal Analysis, Supervision, Writing – Review & Editing; **Bortoluzzi C**: Visualization, Writing – Original Draft Preparation

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. This work was supported by the Genoscope, the Commissariat à l'Énergie Atomique et aux Énergies Alternatives (CEA), France Génomique (ANR-10-INBS-09-08), and the exploratory research programme 'ATLASea: Atlas of marine genomes and its targeted project SEQ-Sea (ANR-22-EXAT-0003-SEQ-Sea).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2025 Verdes A *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License , which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Verdes A, Alvarez-Campos P, Conejero M *et al.* **ERGA-BGE reference genome of the lineid heteronemertean *Lineus lacteus* (Pilidiophora, Nemertea) [version 2; peer review: 2 approved]** Open Research Europe 2025, **5**:317 https://doi.org/10.12688/openreseurope.21501.2

First published: 16 Oct 2025, **5**:317 https://doi.org/10.12688/openreseurope.21501.1

REVISED **Amendments from Version 1**

We here provide an updated version of the Genome Report for Lineus lacteus. In particular, we provide additional details regarding maintenance of the specimens in the lab used for the genome assembly.

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

Introduction

Lineus lacteus, a member of the Lineidae family, is a marine ribbon worm species primarily characterized by its unique morphological and genetic characteristics. The species can reach lengths of up to 60 cm but is only 1–2 mm wide. The body colour is milky white, with a reddish anterior part. The head is elongated and has 6 to 15 eyes on each side arranged in a dorsolateral row. The mouth is located ventrally, far behind the cerebral ganglia, unlike other lineids in which it is much closer (Ament-Velásquez *et al.*, 2016; Bachiller, 2016; Rathke, 1843). The species is distributed throughout the Mediterranean Sea and the North Atlantic Ocean.

Ribbon worms are major predators in marine ecosystems, feeding on species at the top of the food chain, such as predatory annelids and crustaceans, as well as commercially important species like oysters and mussels. Nemerteans use toxins to paralyze their prey and secrete a toxic mucus to defend themselves from predators (Göransson *et al.*, 2019). In addition, many nemerteans have remarkable regenerative capabilities, being able to regrow entire body parts and in some cases, even a whole individual from a small fragment of the body (Zattara *et al.*, 2019).

Developing a high-quality reference genome for *Lineus lacteus* is essential for advancing our understanding of its unique genetic characteristics and adaptive traits. This genomic resource will provide valuable insights into the molecular ecology of an important predator in marine ecosystems and the evolutionary processes that shape venom composition and toxin diversity within the Nemertea, offering broader implications for venomics and toxinology research. In addition, *L. lacteus* belongs to the Lophotrochozoa — a superphylum of animals including leeches, snails and other invertebrates that is crucial to our understanding of bilaterian evolution.

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 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 08 July 2020, Aida Verdes and Patricia Alvarez-Campos sampled eight specimens of *Lineus lacteus* (sex unknown). The specimens were identified by Aida Verdes *in visu* in the Isla de Tabarca, Spain. 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 (https://www.boe.es/eli/rd/2017/24/124). The specimens were preserved alive by Patricia Alvarez-Campos until DNA extraction. Specimen were kept in small tanks with artificial sea water inside an incubator at 20 °C and fed frozen liver pieces weekly.

Vouchering information

Physical reference materials for the sequenced specimen here have been deposited in the Museo Nacional de Ciencias Naturales (MNCN) https://www.mncn.csic.es/en under accession number MNCN 5.03/7.

Frozen reference tissue material from the posterior end of the body is also available from a proxy voucher at the Biobank of the Museo Nacional de Ciencias Naturales (MNCN) https://www.mncn.csic.es/en under voucher ID MNCN-ADN-151756 and from vouchers at the Biobank of the Leibniz Institute for the Analysis of Biodiversity Change https://leibniz-lib.de/en/ under voucher IDs ZFMK-TIS-102878—102938, ZFMK-DNA-FD19596985, ZFMK-RNA-FD19597035—19597036.

Data availability

Lineus lacteus and the related genomic study were assigned to Tree of Life ID (ToLID) 'tnRamLact8' and all sample, sequence, and assembly information are available under the umbrella BioProject PRJEB77793. The sample information is available at the following BioSample accessions: SAMEA114541005, SAMEA114541015, and SAMEA117738716. The genome assembly is accessible from ENA under accession number GCA_965152395.1. The annotated genome will be made 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: ERX13168324, ERX13168323, ERX14048567, and ERX12752257. 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/Lineus_lacteus/tnRamLact8. Further details and data about the project are hosted on the ERGA portal at https://portal.erga-biodiversity.eu/data_portal/947578.

Genetic information

The estimated genome size, based on ancestral taxa, is 1.37 Gb. This is a diploid genome with a haploid number of 16 chromosomes (2n=32). All information for this species was retrieved from Genomes on a Tree (Challis *et al.*, 2023).

DNA/RNA processing

DNA was extracted from mid body using the Blood & Cell Culture DNA Midi 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 used.

RNA was extracted using a RNeasy Mini Kit (Qiagen) according to the manufacturer's instructions. RNA was extracted from two different specimen parts: head and mid body. RNA quantification was performed using the Qubit RNA BR kit and RNA integrity was assessed using a Bioanalyzer 2100 system RNA 6000 Pico Kit (Agilent). The RNA was stored at -80 °C until used.

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 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 posterior body tissue using the Dovetail Omni-C Kit (Cantata Bio), followed by the KAPA Hyper Prep Kit for Illumina sequencing (Roche). The RNA library was prepared from an equimolarly pooled sample using the KAPA mRNA Hyper prep kit (Roche). The short-read libraries were sequenced on a NovaSeq 6000 instrument (Illumina). In total 339x Oxford Nanopore, 274x Illumina WGS shotgun, and 64.6x HiC data were sequenced to generate the assembly.

Genome assembly methods

The genome was assembled using the CNAG CLAWS pipeline (Gomez-Garrido, 2024). Briefly, reads were preprocessed 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 NextDenovo v2.5.0 (Hu *et al.*, 2024), followed by polishing of the assembled contigs using HyPo v1.0.3 (Kundu *et al.*, 2019), removal of retained haplotigs using purge-dups v1.2.6 (Guan *et al.*, 2020) and scaffolding with YaHS v1.2a (Zhou *et al.*, 2023). Finally, assembled scaffolds were curated via manual inspection using Pretext v0.2.5 with the Rapid Curation Toolkit (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). Summary analysis of the released assembly was performed using the ERGA-BGE Genome Report ASM Galaxy workflow (10.48546/workflowhub.workflow.1103.2).

Results

Genome assembly

The genome assembly has a total length of 370,829,418 bp in 19 contiguous chromosomes (Figure 1 & Figure 2), with a GC content of 41.3%. The assembly has a contig N50 of

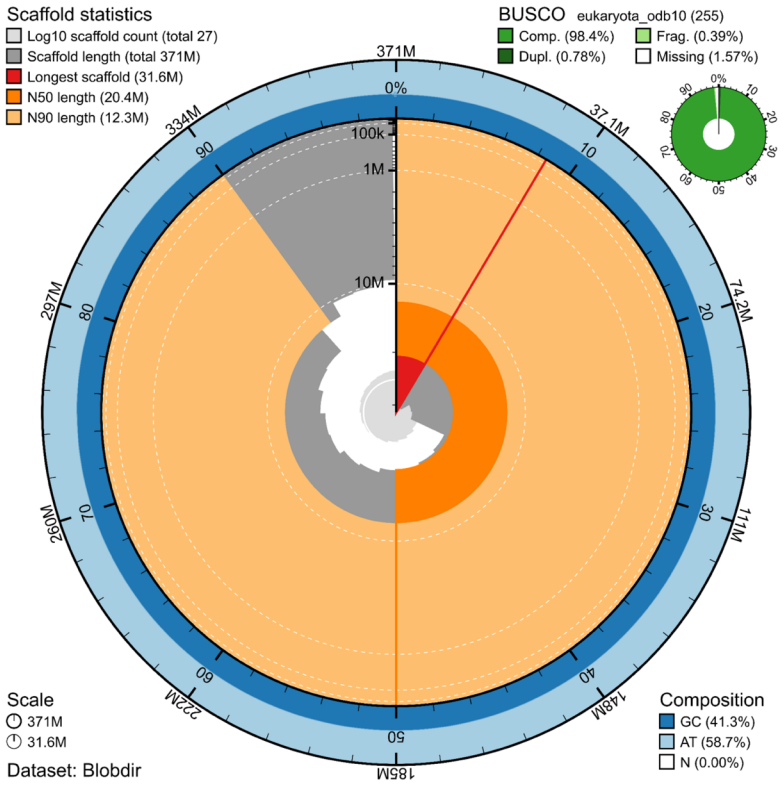


Figure 1. Snail plot summary of assembly statistics. The main plot is divided into 1,000 size-ordered bins around the circumference, with each bin representing 0.1% of the 370,829,418 bp assembly. The distribution of sequence lengths is shown in dark grey, with the plot radius scaled to the longest sequence present in the assembly (31.6 Mb, shown in red). Orange and pale-orange arcs show the scaffold N50 and N90 sequence lengths (20,380,758 and 12,283,121 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.

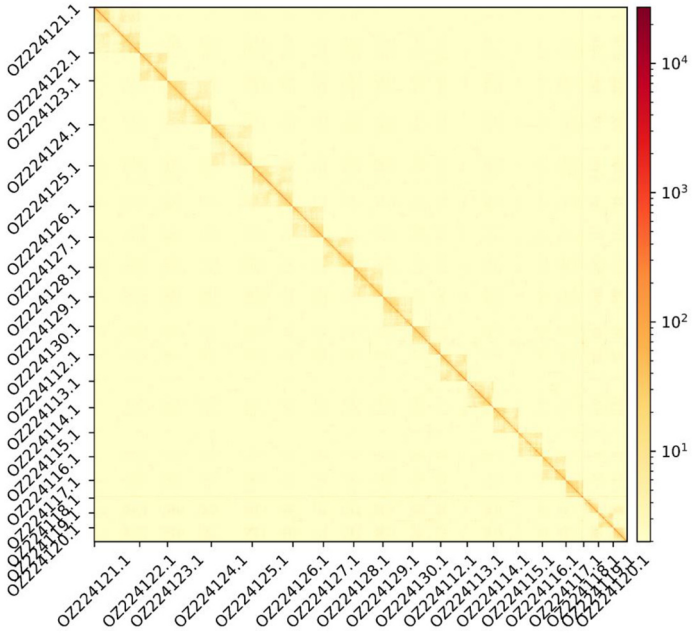


Figure 2. Hi-C contact map showing spatial interactions between regions of the genome. The diagonal corresponds to intra-chromosomal contacts, depicting chromosome boundaries. The frequency of contacts is shown on a logarithmic heatmap scale. Hi-C matrix bins were merged into a 100 kb bin size for plotting.

8,945,886 bp and L50 of 14 and a scaffold N50 of 20,380,758 bp and L50 of 8. The assembly has a total of 44 gaps, totalling 8.8 kb in cumulative size. The single-copy gene content analysis using the Eukaryota database with BUSCO (Manni *et al.*, 2021) resulted in 98.4% completeness (97.6% single and 0.8% duplicated). 76.4% of reads k-mers were present in the assembly and the assembly has a base accuracy Quality Value (QV) of 42.6 as calculated by Merqury (Rhie *et al.*, 2020).

Author contributions

AV and PA-C collected the species; AV identified the species; AV and PA-C sampled and preserved biological material and provided metadata; MC and AR carried out barcoding and managed the deposit of specimens for vouchering and biobanking; AsB, RM, JP-G, and RF provided support in sampling, shipping of biological material, metadata collection, and management; LA and MG extracted DNA, prepared libraries, and performed sequencing; FCF, JGG, and FC performed genome assembly and curation under the supervision of TSA; 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 Juan Junoy from Universidad de Alcalá and Irene del Olmo from Universidad Autónoma de Madrid for their assistance in field sampling and species identification. 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 Jean-Marc Aury from Genoscope.

References

Ament-Velásquez SL, Figuet E, Ballenghien M, *et al.*: **Population genomics of sexual and asexual lineages in fissiparous ribbon worms (Lineus, Nemertea): hybridization, polyploidy and the Meselson effect.** *Mol Ecol.* 2016; **25**(14): 3356–3369.
PubMed Abstract | Publisher Full Text

Bachiller AH: **Los nemertinos de España y Portugal.** [PhD Thesis, Universidad de Alcalá]. 2016.
Reference Source

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**

Gomez-Garrido J: **CLAWS (CNAG's Long-read Assembly Workflow in Snakemake)**. [Computer software], 2024.
Publisher Full Text

Göransson U, Jacobsson E, Strand M, *et al.*: **The toxins of nemertean worms.** *Toxins (Basel).* 2019; **11**(2): 120.
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**

Hu J, Wang Z, Sun Z, *et al.*: **NextDenovo: an efficient error correction and accurate assembly tool for noisy long reads.** *Genome Biol.* 2024; **25**(1): 107.
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.* 2019–12, 2019.
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**(7969): 252.
PubMed Abstract | **Publisher Full Text**

Rathke H: **Beiträge zur fauna Norwegens (Vol. 1)**. für die Akademie in Eduard Weber's Buchhandlng in Bonn, 1843.
Reference Source

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**

Zattara EE, Fernández-Álvarez FA, Hiebert TC, *et al.*: **A phylum-wide survey reveals multiple independent gains of head regeneration in Nemertea.** *Proc Biol Sci.* 2019; **286**(1898): 20182524.
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 19 January 2026

https://doi.org/10.21956/openreseurope.23793.r67370

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

✔ **Natsumi Hookabe**
Japan Agency for Marine-Earth Science and Technology (JAMSTEC), Yokosuka, Japan

The manuscript entitled “ERGA-BGE reference genome of the lineid heteronemertean *Lineus lacteus* (Pilidiophora, Nemertea)” by Verdes et al. is clearly written and presents high-quality genomic data. The methods are sound, the results robust, and the resource valuable for future nemertean and comparative genomic studies. The manuscript is acceptable as submitted.

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: Systematics; Phylogeny; Nemertean biology

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

Author Response 19 Jan 2026
Chiara Bortoluzzi

We thank the reviewer for their positive feedback and for agreeing to the publication of this Genome Report.

Competing Interests: No competing interests were disclosed.

Version 1

Reviewer Report 10 November 2025

https://doi.org/10.21956/openreseurope.23257.r62661

© **2025 Lewin T.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Thomas D Lewin
[1] Biodiversity Research Center Academia Sinica, Taipei City, Taipei City, Taiwan
[2] Biodiversity Research Center Academia Sinica, Taipei City, Taipei City, Taiwan

Verdes et al. present a chromosome-level genome for the nemertean *Lineus lacteus*. Nemerteans are a phylum of vermiform animals within the Lophotrochozoa that is significantly underrepresented among genomic databases compared to related phyla like molluscs and annelids. This new nemertean genome will therefore be of great value to the study of lophotrochozoan genomics and evolution. In particular, the authors highlight the potential for insights into venom production and regeneration given the prevalence of these traits within nemerteans. The genome itself is of high quality and the article is the standard format for a “genome note” style publication. It is well-written and contains all of the necessary information and explanation. The sequencing data and assembled genome are made available for public use. I have no concerns about the quality of the work or the genome and strongly support its indexation.

I have only a couple of very minor comments that the authors may wish to consider.

1. I would expect the genome to have undergone some kind of decontamination procedure as part of the assembly process, such as with BlobToolKit (Challis et al. 2020 G3), but as it stands I can’t see any description of this. Was any decontamination step completed?
2. Currently the manuscript only reports that “the specimens were preserved alive … until DNA extraction”. It might be good to provide a little bit more information here. How long were they kept? What were the culture conditions? Etc etc. Anything that could be useful for someone else looking to do culture these animals.
3. The citation for purge_dups appears to be missing.

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: Evolutionary 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.

Author Response 20 Nov 2025

Chiara Bortoluzzi

In this new version of the Genome Report, we have addressed the reviewer's comments as below:

1. The decontamination step was completed. We did not report these analyses in the main text, because these are already reported in the EAR, which we refer to in the main text using the following link: https://github.com/ERGA-consortium/EARs/tree/main/Assembly_Reports/Lineus_lacteus/tnRamLact8.
2. We provide additional information regarding the maintenance of the specimens in the lab.
3. We added the missing citation for the purge_dups software.

Competing Interests: No competing interests were disclosed.