



DATA NOTE

The chromosomal genome sequence of the tubeworm, *Lamellibrachia columna* Southward, 1991 (Sabellida: Siboglinidae)

[version 1; peer review: 2 approved]

Yanan Sun¹, Jian-Wen Qiu ², Chong Chen ³, José M. Martín-Durán⁴, Graeme Oatley ⁵, Elizabeth Sinclair ⁵, Eerik Aunin⁵, Noah Gettle ⁵, Camilla Santos⁵, Michael Paulini⁵, Haoyu Niu⁵, Victoria McKenna⁵, Rebecca O'Brien ⁵,

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,

EBI Aquatic Symbiosis Genomics Data Portal Team,

Aquatic Symbiosis Genomics Project Leadership

¹Laboratory of Marine Organism Taxonomy and Phylogeny, Institute of Oceanology, Chinese Academy of Sciences, Qingdao, China

²Department of Biology, Hong Kong Baptist University, Hong Kong, China

³X-STAR, Japan Agency for Marine-Earth Science and Technology (JAMSTEC), Yokosuka, Kanagawa, Japan

⁴School of Biological and Behavioural Sciences, Queen Mary University of London, London, England, UK

⁵Tree of Life Programme, Wellcome Sanger Institute, Hinxton, England, UK

V1 First published: 16 Feb 2026, 11:127
<https://doi.org/10.12688/wellcomeopenres.25860.1>
 Latest published: 16 Feb 2026, 11:127
<https://doi.org/10.12688/wellcomeopenres.25860.1>

Abstract

We present a genome assembly from an individual *Lamellibrachia columna* (tubeworm; Annelida; Polychaeta; Sabellida; Siboglinidae). The genome sequence has a total length of 879.73 megabases. Most of the assembly (99.96%) is scaffolded into 15 chromosomal pseudomolecules. The mitochondrial genome has also been assembled, with a length of 16.78 kilobases. Gene annotation of this assembly by Ensembl identified 21 983 protein-coding genes.

Keywords

Lamellibrachia columna; tubeworm; genome sequence; chromosomal; Sabellida; microbial metagenome assembly

Open Peer Review

Approval Status

	1	2
version 1		
16 Feb 2026	view	view
1. Erich M. Schwarz , Cornell University, Ithaca, USA		
2. Thomas D. Lewin , Queen Mary University of London School of Biological and Behavioural Sciences (Ringgold ID: 646267), London, UK		
Any reports and responses or comments on the		



This article is included in the [Tree of Life](#) gateway.

article can be found at the end of the article.

Corresponding author: Aquatic Symbiosis Genomics Project Leadership (mark.blaxter@sanger.ac.uk)

Author roles: Sun Y: Data Curation, Writing – Review & Editing; Qiu JW: Conceptualization, Writing – Original Draft Preparation; Chen C: Investigation, Resources; Martín-Durán JM: Conceptualization, Resources, Writing – Review & Editing; Oatley G: Investigation; Sinclair E: Investigation; Aunin E: Formal Analysis; Gettle N: Formal Analysis; Santos C: Formal Analysis; Paulini M: Formal Analysis; Niu H: Project Administration; McKenna V: Project Administration; O'Brien R: Project Administration;

Competing interests: No competing interests were disclosed.

Grant information: This work was funded by the Gordon and Betty Moore Foundation through a grant (GBMF8897) to the Wellcome Sanger Institute to support the Aquatic Symbiosis Genomics Project, and by Wellcome through core funding to the Wellcome Sanger Institute (220540). The sampling and analysis were also supported by the Shandong Provincial Natural Science Foundation (grant number 2023HWYQ-101), the Taishan Scholars Program (grant number tsqn:202306288) awarded to Yanan Sun, and the Collaborative Research Fund from the University Grants Committee of Hong Kong SAR awarded to Jian-Wen Qiu (grant number C2013-22GF). *The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

Copyright: © 2026 Sun Y *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: Sun Y, Qiu JW, Chen C *et al.* **The chromosomal genome sequence of the tubeworm, *Lamellibrachia columna* Southward, 1991 (Sabellida: Siboglinidae) [version 1; peer review: 2 approved]** Wellcome Open Research 2026, 11:127 <https://doi.org/10.12688/wellcomeopenres.25860.1>

First published: 16 Feb 2026, 11:127 <https://doi.org/10.12688/wellcomeopenres.25860.1>

Species taxonomy

Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Protostomia; Spiralia; Lophotrochozoa; Annelida; Polychaeta; Sedentaria; Canalipalpata; Sabellida; Siboglinidae; *Lamellibrachia*; *Lamellibrachia columna* Southward, 1991 (NCBI:txid53616)

Background

Lamellibrachia columna Southward, 1991 belongs to Vestimentifera, a clade of conspicuous tube-building worms in the annelid family Siboglinidae, inhabiting deep-sea hydrothermal vents, cold-water seeps, and organic falls (Bright & Lallier, 2010). These tubeworms were initially treated as a new order or even phylum due to their highly unusual body plan (Jones, 1985), but were later shown to be a derived clade within Siboglinidae. The body of a vestimentiferan is divided into four sections: the anterior gill plume, the vestimentum, the trunk, and the opisthosoma. Vestimentiferans lack a digestive system and rely on endosymbiotic bacteria in the trophosome of their trunk for energy and nutrients. The bacteria were obtained from the environment during settlement in each generation, and these symbionts utilise hydrogen sulphide contained in geofluids or the sediment as the energy source for chemosynthesis. *Lamellibrachia* is an early-diverging genus of vestimentiferans with nine recognised species. It can be distinguished from other vestimentiferan genera by the following combination of characteristics (Southward, 1991): the gill filaments are parallel to the obturaculum and attached around its base; the obturaculum lacks encrustation or ornamentation; there are sheath lamellae outside the branchial lamellae; and the anterior and posterior ends of the vestimental folds do not fuse.

Southward (1991) described *L. columna* as a new species from the Lau Back-Arc Basin, based on unique features of the shape and size of the tube, the obturaculum, and the number of sheath lamellae, but McCowin *et al.* (2019) commented that only the numbers of sheath and branchial lamellae are distinct features of this species. *L. columna* has been reported from both seep and vent habitats, with a wide distribution from Sagami Bay, Japan, to New Zealand in the Pacific Ocean (McCowin *et al.*, 2019). *Lamellibrachia* worms from Sagami Bay, Okinawa Trough, and Nankai Trough in Japan were initially split into two clades, L1 and L2, based on the mitochondrial cytochrome *c* oxidase subunit I (COI) gene (Kojima *et al.*, 2001), with L1 later named as *L. sagami* Kobayashi *et al.* (2015). However, McCowin *et al.* (2019) later showed that both are conspecific with *L. columna*, based on their morphology and the small genetic distances in two mitochondrial genes (16S, COI) between these populations and specimens from the type locality. The *L. columna* individual used for genome sequencing herein was collected from the Hatsushima seep site (Hashimoto *et al.*, 1989) in Sagami Bay (Figure 1). The tubes of this species measure up to 2 cm in diameter and typically up to 1 m in length, although in extreme cases they may reach up to 2 m.

The *L. columna* reference genome will contribute to understanding of the molecular basis of the metabolic complementarity between annelid hosts and their symbionts and comparison with



Figure 1. Photograph showing a bouquet of *Lamellibrachia columna* at the Off Hatsushima seep, Sagami Bay.

other siboglinid holobionts, such as *Riftia pachyptila* (de Oliveira *et al.*, 2022), *Osedax frankpressi* (Moggioli *et al.*, 2023) and *Paraescarpia echinospica* (Sun *et al.*, 2021), will provide information about the diversity of symbioses that allowed vestimentiferans to colonise various deep-sea chemosynthetic habitats.

Methods

Sample acquisition

The specimen used for genome sequencing was a *Lamellibrachia columna* (specimen ID QMOUL0000077, ToLID wsLam-Colu1; Figure 1), collected from Off Hatsushima seep, Sagami Bay, Japan (latitude 35.0156, longitude 139.2239), at a depth of 929 m on 2021-08-22, during HOV (human-occupied vehicle) *Shinkai 6500* dive 1599 on the R/V *Yokosuka* (cruise YK21-15C). The specimen was collected and identified by Chong Chen (Japan Agency for Marine-Earth Science and Technology). The same specimen was used for RNA sequencing.

Nucleic acid extraction

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 (Howard *et al.*, 2025). The wsLamColu1 sample was weighed and triaged to determine the appropriate extraction protocol. Tissue was homogenised by [powermashing](#) using a PowerMasher II tissue disruptor. HMW DNA was extracted using the [Manual MagAttract](#) protocol. DNA was sheared into an average fragment size of 12–20 kb following the [Megaruptor@3 for LI PacBio](#) protocol. Sheared DNA was purified by [automated SPRI](#) (solid-phase reversible immobilisation). 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. For this sample, the final post-shearing DNA had a Qubit concentration of 18.1 ng/μL and a yield of 2 353.00 ng.

RNA was extracted from thorax tissue of *wsLamColu1* in the Tree of Life Laboratory at the WSI using the [RNA Extraction: Automated MagMax™ mirVana protocol](#). The RNA concentration was assessed using a Nanodrop spectrophotometer and a Qubit Fluorometer using the Qubit RNA Broad-Range Assay kit. Analysis of the integrity of the RNA was done using the Agilent RNA 6000 Pico Kit and Eukaryotic Total RNA assay.

PacBio HiFi 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), following 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 using the Sequel IIe system (Pacific Biosciences, California, USA). The concentration of the library loaded onto the Sequel IIe was in the range 40–135 pM. The SMRT link software, a PacBio web-based end-to-end workflow manager, was used to set-up and monitor the run, and to perform primary and secondary analysis of the data upon completion.

Hi-C

Sample preparation and crosslinking

The Hi-C sample was prepared from 20–50 mg of frozen tissue from the *wsLamColu1* sample using the Arima-HiC v2 kit (Arima Genomics). Following the manufacturer's instructions, tissue was fixed and DNA crosslinked using TC buffer to a final formaldehyde concentration of 2%. The tissue was homogenised using the Diagnocine Power Masher-II. Crosslinked DNA was digested with a restriction enzyme master mix, biotinylated, and ligated. Clean-up was performed with SPRISelect beads before library preparation. DNA concentration was measured with the Qubit Fluorometer (Thermo Fisher Scientific) and Qubit HS Assay Kit. The biotinylation percentage was estimated using the Arima-HiC v2 QC beads.

Hi-C library preparation and sequencing

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 to create 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 6000.

RNA library preparation and sequencing

Libraries were prepared using the NEBNext® Ultra™ II Directional RNA Library Prep Kit for Illumina (New England Biolabs), following the manufacturer's instructions. Poly(A) mRNA in the total RNA solution was isolated using oligo(dT) beads, converted to cDNA, and uniquely indexed; 14 PCR cycles were performed. Libraries were size-selected to produce fragments between 100–300 bp. Libraries were quantified, normalised, pooled to a final concentration of 2.8 nM, and diluted to 150 pM for loading. Sequencing was carried out on the Illumina NovaSeq 6000, generating paired-end reads.

Genome assembly

Prior to assembly of the PacBio HiFi reads, a database of *k*-mer counts (*k* = 31) was generated from the filtered reads using [FastK](#). GenomeScope2 ([Ranallo-Benavidez et al., 2020](#)) was used to analyse the *k*-mer frequency distributions, providing estimates of genome size, heterozygosity, and repeat content.

The HiFi reads were assembled using Hifiasm ([Cheng et al., 2021](#)) with the `--primary` option. Haplotypic duplications were identified and removed using `purge_dups` ([Guan et al., 2020](#)). The Hi-C reads ([Rao et al., 2014](#)) were mapped to the primary contigs using `bwa-mem2` ([Vasimuddin et al., 2019](#)), and the contigs were scaffolded in YaHS ([Zhou et al., 2023](#)) with the `--break` option for handling potential misassemblies. The 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 MitoHiFi ([Uliano-Silva et al., 2023](#)).

Assembly curation

The assembly was decontaminated using the Assembly Screen for Cobionts and Contaminants (ASCC) pipeline. [TreeVal](#) was used to generate the flat files and maps for use in curation. Manual curation was conducted primarily in [PretextView](#) and HiGlass ([Kerpedjiev et al., 2018](#)). Scaffolds were visually inspected and corrected as described by [Howe et al. \(2021\)](#). Manual corrections included 72 breaks, 83 joins, and removal of 55 haplotypic duplications. This reduced the scaffold count by 54.7%, increased the scaffold N50 by 13.1%, and reduced the

total assembly length by 2.6%. The curation process is described at <https://gitlab.com/wtsi-grit/rapid-curation>. PretextSnapshot was used to generate a Hi-C contact map of the final assembly.

Assembly quality assessment

The Mercury.FK tool (Rhie *et al.*, 2020) was run in a Singularity container (Kurtzer *et al.*, 2017) to evaluate k -mer completeness and assembly quality for the primary and alternate haplotypes using the k -mer databases ($k = 31$) computed prior to genome assembly. The analysis outputs included assembly QV scores and completeness statistics.

The genome was analysed using the BlobToolKit pipeline, a Nextflow implementation of the earlier Snakemake version (Challis *et al.*, 2020). The pipeline aligns PacBio reads using minimap2 (Li, 2018) and SAMtools (Danecek *et al.*, 2021) to generate coverage tracks. It runs BUSCO (Manni *et al.*, 2021) using lineages identified from the NCBI Taxonomy (Schoch *et al.*, 2020). For the three domain-level lineages, BUSCO genes are aligned to the UniProt Reference Proteomes database (Bateman *et al.*, 2023) using DIAMOND blastp (Buchfink *et al.*, 2021). The genome is divided into chunks based on the density of BUSCO genes from the closest taxonomic lineage, and each chunk is aligned to the UniProt Reference Proteomes database with DIAMOND blastx. Sequences without hits are chunked using seqtk and aligned to the NT database with blastn (Altschul *et al.*, 1990). The BlobToolKit suite

consolidates all outputs into a blobdir for visualisation. The BlobToolKit pipeline was developed using nf-core tooling (Ewels *et al.*, 2020) and MultiQC (Ewels *et al.*, 2016), with containerisation through Docker (Merkel, 2014) and Singularity (Kurtzer *et al.*, 2017).

Genome sequence report

Sequence data

PacBio sequencing of the *Lamellibrachia columna* specimen generated 66.46 Gb (gigabases) from 7.58 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 1 026.13 Mb, with a heterozygosity of 1.59% and repeat content of 64.14% (Figure 2). These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 38× coverage. Hi-C sequencing produced 231.43 Gb from 1 532.68 million reads, which were used to scaffold the assembly. RNA sequencing data were also generated and are available in public sequence repositories. Table 1 summarises the specimen and sequencing details.

Assembly statistics

The primary haplotype was assembled, and contigs corresponding to an alternate haplotype were also deposited in INSDC databases. The final assembly has a total length of 879.73 Mb in 42 scaffolds, with 706 gaps, and a scaffold N50 of 59.64 Mb (Table 2).

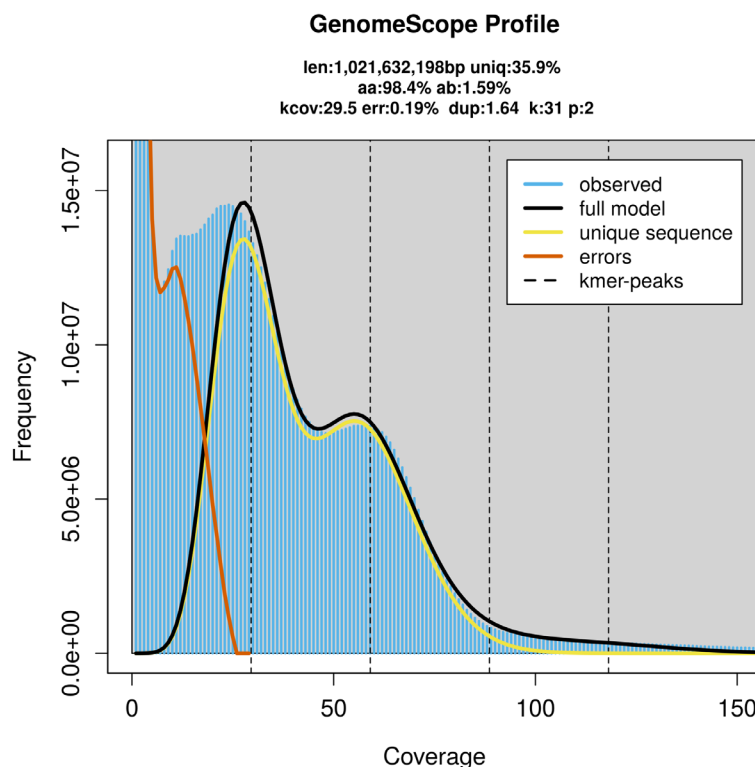


Figure 2. Frequency distribution of k -mers generated using GenomeScope2. The plot shows observed and modelled k -mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

Table 1. Specimen and sequencing data for BioProject PRJEB68010.

Platform	PacBio HiFi	Hi-C	RNA-seq
ToLID	wsLamColu1	wsLamColu1	wsLamColu1
Specimen ID	QMOUL0000077	QMOUL0000077	QMOUL0000077
BioSample (source individual)	SAMEA12097778	SAMEA12097778	SAMEA12097778
BioSample (tissue)	SAMEA12097838	SAMEA12097839	SAMEA12097839
Instrument	Sequel IIe	Illumina NovaSeq 6000	Illumina NovaSeq 6000
Run accessions	ERR12205280; ERR12205278; ERR12205279; ERR14749934	ERR12245605; ERR12245606	ERR12245607
Read count total	7.58 million	1 532.68 million	68.02 million
Base count total	66.46 Gb	231.43 Gb	10.27 Gb

Table 2. Genome assembly statistics.

Assembly name	wsLamColu1.1
Assembly accession	GCA_963662155.1
Alternate haplotype accession	GCA_963667965.1
Assembly level	chromosome
Span (Mb)	879.73
Number of chromosomes	15
Number of contigs	748
Contig N50	1.93 Mb
Number of scaffolds	42
Scaffold N50	59.64 Mb
Organelles	Mitochondrion: 16.78 kb

Most of the assembly sequence (99.96%) was assigned to 15 chromosomal-level scaffolds. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 3; Table 3).

The mitochondrial genome was also assembled (length 16.78 kb, OY759276.1). This sequence is included as a contig in the multifasta file of the genome submission and as a standalone record.

Assembly quality metrics

The combined primary and alternate assemblies achieve an estimated QV of 54.0. The *k*-mer completeness is 76.92% for the primary assembly, 75.61% for the alternate haplotype, and 94.12% for the combined assemblies (Figure 4).

BUSCO v.5.5.0 analysis using the reference set (*n* =) identified % of the expected gene set (single = %, duplicated = %). The

snail plot in Figure 5 summarises the scaffold length distribution and other assembly statistics for the primary assembly. The blob plot in Figure 6 shows the distribution of scaffolds by GC proportion and coverage.

Table 4 lists the assembly metric benchmarks adapted from Rhie *et al.* (2021) and the Earth BioGenome Project Report on Assembly Standards September 2024. The EBP metric, calculated for the primary assembly, is **6.C.Q53**, meeting the recommended reference standard.

Genome annotation report

The *Lamellibrachia columna* genome assembly (GCA_963662155.1) was annotated by Ensembl at the European Bioinformatics Institute (EBI). This annotation includes 36 475 transcribed mRNAs from 21 983 protein-coding and 4 157 non-coding genes. The average transcript length is 14 550.97 bp, with an average of 1.40 coding transcripts per gene and 7.12

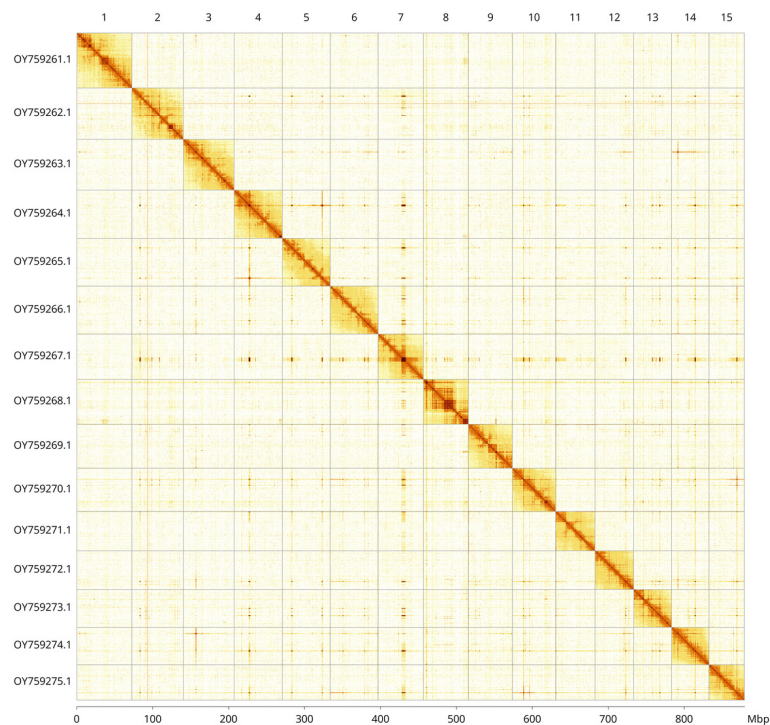


Figure 3. Hi-C contact map of the *Lamellibrachia columnaris* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot. An interactive version of the contact map may be viewed in [HiGlass](#).

Table 3. Chromosomal pseudomolecules in the primary genome assembly of *Lamellibrachia columnaris* wsLamColu1.

INSDC accession	Molecule	Length (Mb)	GC%
OY759261.1	1	72.70	40.50
OY759262.1	2	67.76	40.50
OY759263.1	3	66.82	40.50
OY759264.1	4	63.57	40.50
OY759265.1	5	63.17	40.50
OY759266.1	6	63.05	40.50
OY759267.1	7	59.64	40.50
OY759268.1	8	59.20	40.50
OY759269.1	9	58.07	40.50
OY759270.1	10	56.89	40.50
OY759271.1	11	51.82	40.50
OY759272.1	12	50.84	40.50
OY759273.1	13	49.93	40.50
OY759274.1	14	49.26	40.50
OY759275.1	15	46.62	40.50

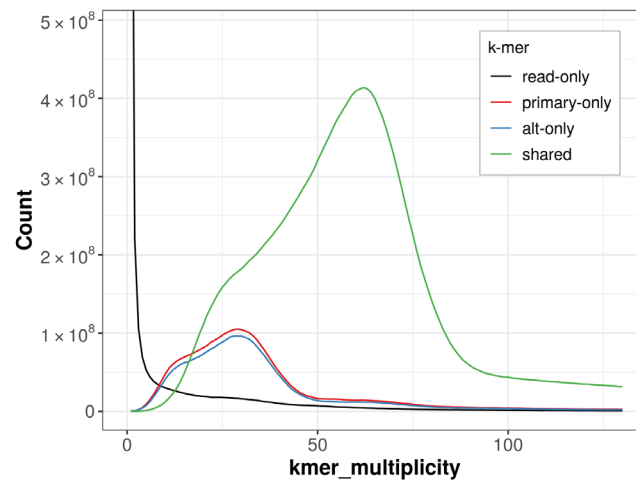


Figure 4. Evaluation of *k*-mer completeness using MerquyFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers that appear in the reads but are not assembled. The green curve corresponds to *k*-mers shared by both haplotypes, and the red and blue curves show *k*-mers found only in one of the haplotypes.

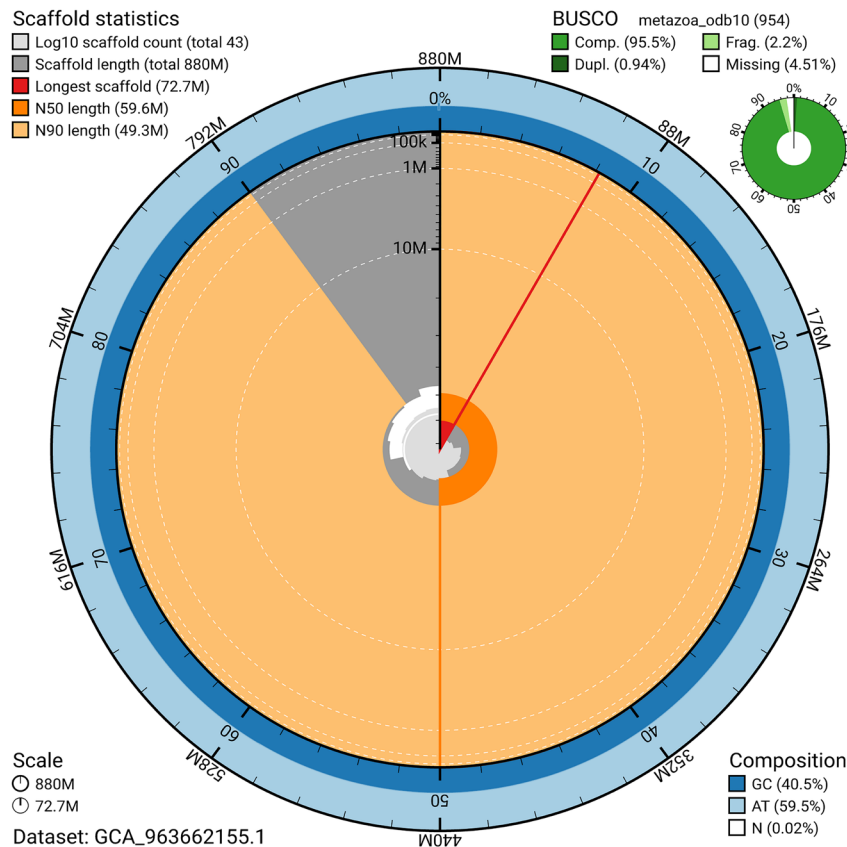


Figure 5. Assembly metrics for wsLamColu1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1 000 bins around the circumference. The outermost blue tracks display the distribution of GC, AT, and N percentages across the bins. Scaffolds are arranged clockwise from longest to shortest and are depicted in dark grey. The longest scaffold is indicated by the red arc, and the deeper orange and pale orange arcs represent the N50 and N90 lengths. A light grey spiral at the centre shows the cumulative scaffold count on a logarithmic scale. A summary of complete, fragmented, duplicated, and missing BUSCO genes in the set is presented at the top right. An interactive version of this figure can be accessed on the [BlobToolKit viewer](#).

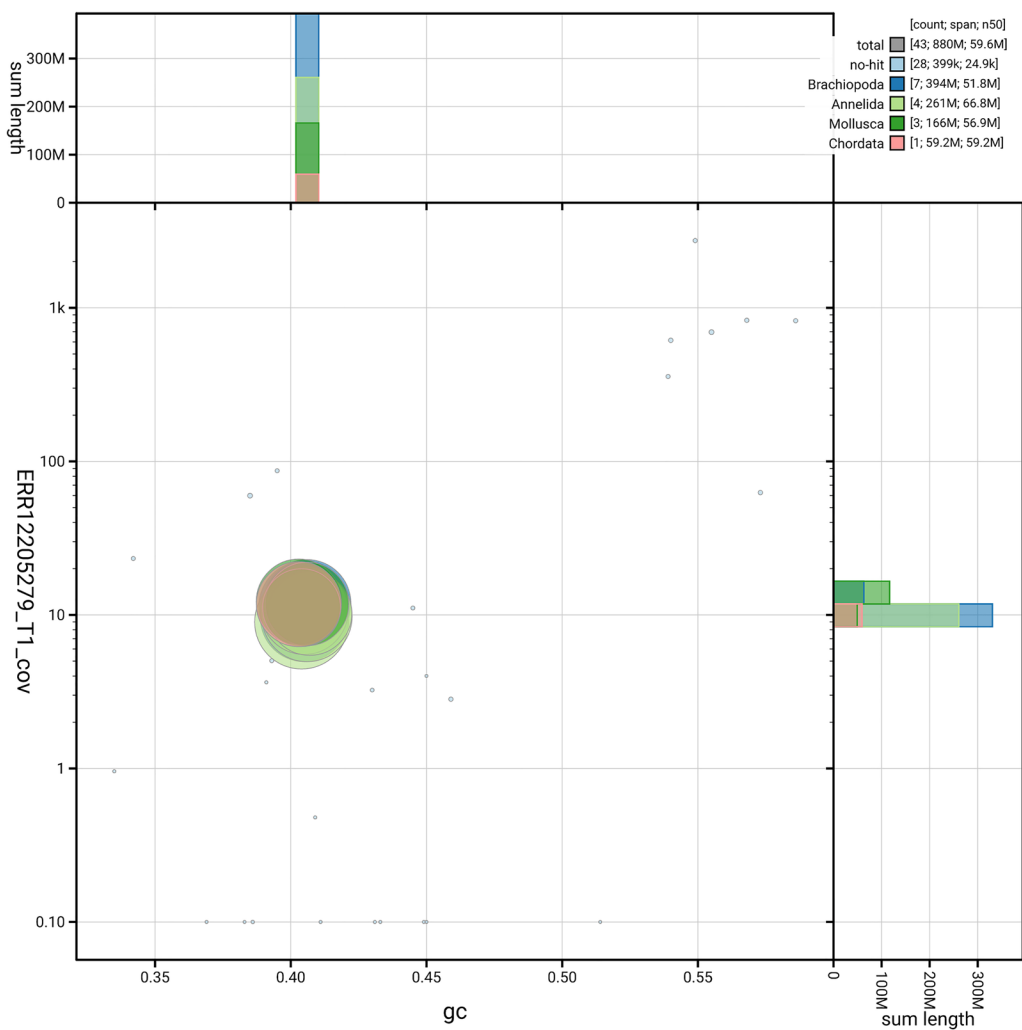


Figure 6. BlobToolKit GC-coverage plot for wsLamColu1.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes display the total length of sequences distributed across different levels of coverage and GC content. An interactive version of this figure is available on the [BlobToolKit viewer](#).

Table 4. Earth Biogenome Project summary metrics for the *Lamellibrachia columna* assembly.

Measure	Value	Benchmark
EBP summary (primary)	6.C.Q53	6.C.Q40
Contig N50 length	1.93 Mb	≥ 1 Mb
Scaffold N50 length	59.64 Mb	= chromosome N50
Consensus quality (QV)	Primary: 53.7; alternate: 54.4; combined: 54.0	≥ 40
<i>k</i> -mer completeness	Primary: 76.92%; alternate: 75.61%; combined: 94.12%	≥ 95%
BUSCO	C:94.6%[S:93.8%,D:0.8%];F:3.0%,M:2.4%,n:954	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	99.96%	≥ 90%

Notes: EBP summary uses log10(Contig N50); chromosome-level (C) or log10(Scaffold N50); Q (Merqurey QV). BUSCO: C=complete; S=single-copy; D=duplicated; F=fragmented; M=missing; n=orthologues

exons per transcript. For further information about the annotation, please refer to the [Ensembl annotation page](#).

Wellcome Sanger Institute – Legal and Governance

The materials that have contributed to this genome note have been supplied by a Tree of Life collaborator. The Wellcome Sanger Institute employs a process whereby due diligence is carried out proportionate to the nature of the materials themselves, and the circumstances under which they have been/are to be collected and provided for use. The purpose of this is to address and mitigate any potential legal and/or ethical implications of receipt and use of the materials as part of the research project, and to ensure that in doing so we align with best practice wherever possible.

The overarching areas of consideration are:

- Ethical review of provenance and sourcing of the material
- Legality of collection, transfer and use (national and international)

Each transfer of samples is undertaken according to a Research Collaboration Agreement or Material Transfer Agreement entered into by the Tree of Life collaborator, Genome Research Limited (operating as the Wellcome Sanger Institute) and in some circumstances other Tree of Life collaborators.

Data availability

European Nucleotide Archive: *Lamellibrachia columna*. Accession number [PRJEB68010](#). The genome sequence is released

openly for reuse. The *Lamellibrachia columna* genome sequencing initiative is part of the Aquatic Symbiosis Genomics Project (PRJEB43743) and Sanger Institute Tree of Life Programme (PRJEB43745). All raw sequence data and the assembly have been deposited in INSDC databases. Raw data and assembly accession identifiers are reported in [Table 1](#) and [Table 2](#).

Production code used in genome assembly at the WSI Tree of Life is available at <https://github.com/sanger-tol>. [Table 5](#) lists software versions used in this study.

Author information

Contributors are listed at the following links:

- Members of the [Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team](#)
- Members of [Wellcome Sanger Institute Scientific Operations – Sequencing Operations](#)
- Members of the [Wellcome Sanger Institute Tree of Life Core Informatics team](#)
- Members of the [EBI Aquatic Symbiosis Genomics Data Portal Team](#)
- The [Aquatic Symbiosis Genomics Project leadership](#)

Acknowledgements

We thank the captain and crew of R/V *Yokosuka* as well as the HOV *Shinkai 6500* for their support during cruise YK23-15C, Ken Takai (JAMSTEC) for leading the research expedition, and Pei-Yuan Qian (The Hong Kong University of Science and Technology) for co-organising the cruise.

Table 5. Software versions and sources.

Software	Version	Source
BEDTools	2.30.0	https://github.com/arq5x/bedtools2
BLAST	2.14.0	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/
BlobToolKit	4.3.3	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.5.0	https://gitlab.com/ezlab/busco
bwa-mem2	2.2.1	https://github.com/bwa-mem2/bwa-mem2
Cooler	0.8.11	https://github.com/open2c/cooler
DIAMOND	2.1.8	https://github.com/bbuchfink/diamond
fasta_windows	0.2.4	https://github.com/tolkit/fasta_windows
FastK	1.1	https://github.com/thegenemyers/FASTK
Gfastats	1.3.6	https://github.com/vgl-hub/gfastats
GenomeScope2.0	2.0.1	https://github.com/tbenavi1/genomescope2.0
Hifiasm	0.19.5-r587	https://github.com/chhy1p123/hifiasm
HiGlass	1.13.4	https://github.com/higlass/higlass

Software	Version	Source
MercuryFK	1.1.2	https://github.com/thegenemyers/MERQUERY.FK
Minimap2	2.24-r1122	https://github.com/lh3/minimap2
MitoHiFi	3	https://github.com/marcelauliano/MitoHiFi
MultiQC	1.14; 1.17 and 1.18	https://github.com/MultiQC/MultiQC
Nextflow	23.04.1	https://github.com/nextflow-io/nextflow
PretextSnapshot	0.0.5	https://github.com/sanger-tol/PretextSnapshot
PretextView	1.0.3	https://github.com/sanger-tol/PretextView
Seqtk	1.3	https://github.com/lh3/seqtk
Singularity	3.9.0	https://github.com/sylabs/singularity
sanger-tol/ascc	0.1.0	https://github.com/sanger-tol/ascc
sanger-tol/blobtoolkit	0.3.0	https://github.com/sanger-tol/blobtoolkit
sanger-tol/curationpretext	1.4.2	https://github.com/sanger-tol/curationpretext
TreeVal	1.4.0	https://github.com/sanger-tol/treeval
YaHS	1.2a.2	https://github.com/c-zhou/yahs

References

- Altschul SF, Gish W, Miller W, et al.: **Basic Local Alignment Search Tool.** *J Mol Biol.* 1990; **215**(3): 403–410.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Bateman A, Martin MJ, Orchard S, et al.: **UniProt: the Universal Protein knowledgebase in 2023.** *Nucleic Acids Res.* 2023; **51**(D1): D523–D531.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Bright M, Lallier FH: **The biology of vestimentiferan tubeworms.** *Oceanography and Marine Biology: An Annual Review.* 2010; **48**: 213–266.
[Publisher Full Text](#)
- Buchfink B, Reuter K, Drost HG: **Sensitive protein alignments at Tree-of-Life scale using DIAMOND.** *Nat Methods.* 2021; **18**(4): 366–368.
[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](#)
- 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](#)
- Danecek P, Bonfield JK, Liddle J, et al.: **Twelve years of SAMtools and BCFtools.** *GigaScience.* 2021; **10**(2): giab008.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- de Oliveira AL, Mitchell J, Girguis P, et al.: **Novel insights on obligate symbiont lifestyle and adaptation to chemosynthetic environment as revealed by the giant tubeworm genome.** *Mol Biol Evol.* 2022; **39**(1): msab347.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Ewels P, Magnusson M, Lundin S, et al.: **MultiQC: summarize analysis results for multiple tools and samples in a single report.** *Bioinformatics.* 2016; **32**(19): 3047–3048.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Ewels PA, Peltzer A, Fillinger S, et al.: **The nf-core framework for community-curated bioinformatics pipelines.** *Nat Biotechnol.* 2020; **38**(3): 276–278.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Formenti G, Abueg L, Brajuka A, et al.: **Gfstats: conversion, evaluation and manipulation of genome sequences using assembly graphs.** *Bioinformatics.* 2022; **38**(17): 4214–4216.
[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](#)
- Hashimoto J, Ohta S, Tanaka T, et al.: **Deep-sea communities dominated by the giant clam, *Calyptogena soyoe*, along the slope foot of Hatsushima Island, Sagami Bay Central Japan.** *Palaeogeogr Palaeoclimatol Palaeoecol.* 1989; **71**(1–2): 179–192.
[Publisher Full Text](#)
- 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.
[Publisher Full Text](#)
- Howe K, Chow W, Collins J, et al.: **Significantly improving the quality of genome assemblies through curation.** *GigaScience.* 2021; **10**(1): g1aa153.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Jones ML: **On the vestimentifera, new phylum: six new species, and other taxa, from hydrothermal vents and elsewhere.** *Bull Biol Soc Wash.* 1985; **6**: 117–158.
[Reference Source](#)
- 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](#)
- Kobayashi G, Miura T, Kojima S: ***Lamellibrachia sagami* sp. nov., a new vestimentiferan tubeworm (Annelida: Siboglinidae) from Sagami Bay and several sites in the northwestern Pacific Ocean.** *Zootaxa.* 2015; **4018**(1): 97–108.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Kojima S, Ohta S, Yamamoto T, et al.: **Molecular taxonomy of vestimentiferans of the Western Pacific and their phylogenetic relationship to species of the Eastern Pacific. I. Family Lamellibrachidae.** *Mar Biol.* 2001; **139**: 211–219.
[Publisher Full Text](#)

Kurtzer GM, Sochat V, Bauer MW: **Singularity: scientific containers for mobility of compute.** *PLoS One.* 2017; **12**(5): e0177459.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Li H: **Minimap2: pairwise alignment for nucleotide sequences.** *Bioinformatics.* 2018; **34**(18): 3094–3100.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Manni M, Berkeley MR, Seppely 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): 4647–4054.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

McCowin MF, Rowden AA, Rouse GW: **A new record of *Lamellibrachia columna* (Siboglinidae, Annelida) from cold seeps off New Zealand, and an assessment of its presence in the western Pacific Ocean.** *Mar Biodivers Rec.* 2019; **12**(1): 10.
[Publisher Full Text](#)

Merkel D: **Docker: lightweight Linux containers for consistent development and deployment.** *Linux J.* 2014; **2014**(239): 2.
[Reference Source](#)

Moggioli G, Panossian B, Sun Y, *et al.*: **Distinct genomic routes underlie transitions to specialised symbiotic lifestyles in deep-sea annelid worms.** *Nat Commun.* 2023; **14**: 2814.
[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](#)

Rao SSP, Huntley MH, Durand NC, *et al.*: **A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping.** *Cell.* 2014; **159**(7): 1665–1680.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Rhie A, McCarthy SA, Fedrigo O, *et al.*: **Towards complete and error-free genome assemblies of all vertebrate species.** *Nature.* 2021; **592**(7856): 737–746.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

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

Schoch CL, Ciufo S, Domrachev M, *et al.*: **NCBI taxonomy: a comprehensive update on curation, resources and tools.** *Database (Oxford).* 2020; **2020**: baaa062. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Southward EC: **Three new species of Pogonophora, including two vestimentiferans, from hydrothermal sites in the Lau Back-Arc Basin (Southwest Pacific Ocean).** *J Nat Hist.* 1991; **25**(4): 859–881.
[Publisher Full Text](#)

Sun Y, Sun J, Yang Y, *et al.*: **Genomic signatures supporting the symbiosis and formation of chitinous tube in the deep-sea tubeworm *Paraescarpia echinospica*.** *Mol Biol Evol.* 2021; **38**(10): 4116–4634.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Uliano-Silva M, Ferreira JGRN, Krashenninnikova K, *et al.*: **MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads.** *BMC Bioinformatics.* 2023; **24**(1): 288.
[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.** In: *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. IEEE, 2019; 314–324.
[Publisher 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 1

Reviewer Report 14 April 2026

<https://doi.org/10.21956/wellcomeopenres.28463.r152196>

© 2026 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 

Queen Mary University of London School of Biological and Behavioural Sciences (Ringgold ID: 646267), London, England, UK

Sun et al. present a reference assembly for the tubeworm *Lamellibrachia columna*. As a deep-sea annelid with chemosynthetic endosymbiotic bacteria, this is a key species for those studying symbiosis, deep-sea environments, and annelid evolution.

The note is of the standard format produced by the Tree of Life Programme. Methods and protocols are very well documented and the genome is extremely high quality. It also has a more extensive introduction to the species than many other genome notes which gives a nice pointer to why this particular species is interesting and how the genome could be of use. I have no major concerns about the manuscript and only a couple of minor queries.

First, regarding the BlobToolKit output in Figure 6: if I am reading it correctly, only four of the 43 scaffolds have a phylum membership of Annelida while 7 are Brachiopoda, 3 Mollusca, and 28 have no hit. I expect this kind of result with very understudied phyla that may be unrepresented in databases but I am quite surprised to see it for an annelid. Do the authors have an explanation for this result?

Second, could the authors expand on very specific terms like vestimentum and opisthosoma to make the background more accessible to those who are interested in the genome but not in the annelid field.

Finally, I agree with the remarks of the first reviewer who noted:

1. The absence of BUSCO data in the text and a discrepancy in the BUSCO results reported in Table 4 and Figure 5
2. The absence of an explanation of the genome annotation procedure. I have raised this issue regarding previous genome notes and think it is an important point to be corrected.

Overall, this is a high-quality genome that will be of significant benefit to the field.

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 of spiralian

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 06 April 2026

<https://doi.org/10.21956/wellcomeopenres.28463.r152199>

© 2026 Schwarz E. 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.



Erich M. Schwarz 

Cornell University, Ithaca, New York, USA

Sun et al. provide a chromosomally scaffolded genome sequence for *Lamellibrachia columna*. Their report is generally satisfactory. However, I do have a few corrections to suggest.

In the Grant Information section: "(grant number 2023HWYQ-101), the Taishan Scholars Program" might be better written as "(grant number 2023HWYQ-101) and the Taishan Scholars Program".

In the legend of Figure 1, "Samagi Bay" appears to be a typographical error; it should probably read "Sagami Bay".

In the 'Assembly quality metrics' section of the Results, the BUSCO results need to be properly filled in. Presently, they read: "BUSCO v.5.5.0 analysis using the [which?] reference set (n =) identified % of the expected gene set (single = %, duplicated = %)." If I were to try to correct this from the numbers given in Figure 5, then I would rewrite this as: "BUSCO v.5.5.0 analysis using the metazoa_odb10 reference set (n = 954) identified 96.4% of the expected gene set (single = 95.5%, duplicated = 0.94%)." However, these numbers have another problem; they are inconsistent with

the BUSCO values given in Table 4! The authors should review the BUSCO values given in Figure 5 versus Table 4, explain any apparent discrepancies, correct any real discrepancies, and fill in the sentence in 'Assembly quality metrics' appropriately.

In the Genome Annotation Report, the authors write: "The *Lamellibrachia columna* genome assembly (GCA_963662155.1) was annotated by Ensembl at the European Bioinformatics Institute (EBI)." They later refer readers to the Ensembl web page <https://beta.ensembl.org/species/0d55dde9-c871-4112-88c2-3e8491b8d41d> for details. However, this web page provides no details about *how* these annotations were generated. There must be, somewhere, some reference or web manual by Ensembl that explains in detail how these gene predictions were made by Ensembl. Could the authors please provide a citation to these detailed prediction methods?

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?

Partly

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

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Nematode genome/transcriptome analysis

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.
