Check for updates

DATA NOTE

The genome sequence of the Atlantic Strawberry Cockle, *Americardia media* (Linnaeus, 1758) (Cardiida: Cardiidae)

[version 1; peer review: 2 approved]

Ruiqi Li[1-3], Jingchun Li[1,2],
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,
Tree of Life Core Informatics collective, Mark L. Blaxter[4]

[1]Department of Ecology and Evolutionary Biology, University of Colorado Boulder, Boulder, Colorado, USA
[2]Museum of Natural History, University of Colorado Boulder, Boulder, Colorado, USA
[3]Department of Biological Sciences, University of Southern California, Los Angeles, California, USA
[4]Tree of Life, Wellcome Sanger Institute, Hinxton, England, UK

v1 **First published:** 06 Nov 2025, **10**:630
https://doi.org/10.12688/wellcomeopenres.25082.1
Latest published: 06 Nov 2025, **10**:630
https://doi.org/10.12688/wellcomeopenres.25082.1

Abstract

We present a genome assembly from an individual *Americardia media* (Atlantic Strawberry Cockle; Mollusca; Bivalvia; Cardiida; Cardiidae). The assembly contains two haplotypes with total lengths of 1 299.87 megabases and 1 284.99 megabases. Most of haplotype 1 (99.42%) is scaffolded into 19 chromosomal pseudomolecules. Haplotype 2 was assembled to scaffold level. The mitochondrial genome has also been assembled, with a length of 47.2 kilobases.

Keywords

Americardia media; Atlantic Strawberry Cockle; genome sequence; chromosomal; Cardiida

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ✓

	1	2
version 1 06 Nov 2025	✓ view	✓ view

1. **Kyle E McElroy**, Iowa State University, Ames, USA

2. **Rebekah L. Rogers**, UNC Charlotte, Charlotte, USA

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

Corresponding author: Mark L. Blaxter (mark.blaxter@sanger.ac.uk)

Author roles: Li R: Investigation, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Li J**: Investigation, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Blaxter ML**: Supervision

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome through core funding to the Wellcome Sanger Institute (220540).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2025 Li R *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: Li R, Li J, Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team *et al.* **The genome sequence of the Atlantic Strawberry Cockle, *Americardia media* (Linnaeus, 1758) (Cardiida: Cardiidae) [version 1; peer review: 2 approved]** Wellcome Open Research 2025, **10**:630 https://doi.org/10.12688/wellcomeopenres.25082.1

First published: 06 Nov 2025, **10**:630 https://doi.org/10.12688/wellcomeopenres.25082.1

Species taxonomy

Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Protostomia; Spiralia; Lophotrochozoa; Mollusca; Bivalvia; Autobranchia; Heteroconchia; Euheterodonta; Imparidentia; Neoheterodontei; Cardiida; Cardioidea; Cardiidae; Fraginae; *Americardia*; *Americardia media* (Linnaeus, 1758) (NCBI: txid241156)

Background

Americardia media is a widespread and morphologically variable cockle of the Caribbean and western Atlantic. Its range extends from southern Florida and the Gulf of Mexico southward to Panama, Suriname, and throughout the West Indies (ter Poorten, 2024). This species inhabits sandy bottoms in shallow, sheltered environments, typically in intertidal and subtidal zones associated with seagrass beds or algal meadows (ter Poorten, 2024).

Americardia media belongs to the subfamily Fraginae (heart cockles), which includes both a non-symbiotic lineage and a symbiotic lineage that maintains photosymbiotic relationships with dinoflagellates of the family Symbiodiniaceae (Kirkendale, 2009; Li *et al.*, 2020). This diversity makes Fraginae a powerful system for comparative studies on the origins and molecular mechanisms of animal photosymbiosis. While several genomes from photosymbiotic fraginids are already available (Li *et al.*, 2024a; Li *et al.*, 2024b; Li *et al.*, 2024c), the genome of *A. media* adds an essential non-symbiotic counterpart, enabling investigations into morphological adaptation to photosymbiosis and the divergent evolutionary trajectories of symbiotic versus non-symbiotic bivalves.

We present a chromosome-level genome sequence for *Americardia media*, the first genome for the genus *Americardia* and one of 14 genomes available for the family Cardiidae as of September 2025 (data obtained via NCBI datasets, O'Leary *et al.*, 2024). The assembly was produced using the Tree of Life pipeline from a specimen collected in West Palm Beach, Florida, Usa (Figure 1)

Methods

Sample acquisition

The specimen used for genome sequencing was an adult *Americardia media* (specimen ID SAN20002327, ToLID xbAmeMedi1; Figure 1), collected from West Palm Beach, Florida, USA (latitude 26.783, longitude –80.043) on 2022-06-01. The specimen was collected and identified by Ruiqi Li (University of Colorado, Boulder). The voucher id is UCM51385, deposited in the University of Colorado Boulder Museum of Natural History.

Figure 1. Photograph of the *Americardia media* (xbAmeMedi1) specimen used for genome 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 xbAmeMedi1 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 Automated MagAttract v2 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 4.4 ng/μL and a yield of 572.00 ng.

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

Hi-C

Sample preparation and crosslinking

The Hi-C sample was prepared from 20–50 mg of frozen tissue of the xbAmeMedi1 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 to 16 PCR cycles. Post-PCR clean-up was performed with SPRISelect beads. Libraries were quantified using the AccuClear Ultra High Sensitivity dsDNA Standards Assay Kit (Biotium) and a FLUOstar Omega plate reader (BMG Labtech).

Prior to sequencing, libraries were normalised to 10 ng/μL. Normalised libraries were quantified again and equimolar and/or weighted 2.8 nM pools were created. Pool concentrations were checked using the Agilent 4200 TapeStation (Agilent) with High Sensitivity D500 reagents before sequencing. Sequencing was performed using paired-end 150 bp reads on the Illumina NovaSeq X.

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 in Hi-C phasing mode (Cheng *et al.*, 2021; Cheng *et al.*, 2022), producing two haplotypes. Hi-C reads (Rao *et al.*, 2014) were mapped to the primary contigs using bwa-mem2 (Vasimuddin *et al.*, 2019). Contigs were further scaffolded with Hi-C data in YaHS (Zhou *et al.*, 2023), using 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), which runs MitoFinder (Allio *et al.*, 2020) and uses these annotations to select the final mitochondrial contig and to ensure the general quality of the sequence. The mitochondrial genome was assembled using OATK (Zhou *et al.*, 2025).

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 21 breaks and 44 joins. This reduced the scaffold count by 49.6% and reduced the total assembly length by 1.8%. 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 Merqury.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 both 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 *Americardia media* specimen generated 67.68 Gb (gigabases) from 6.66 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 1 251.78 Mb, with a heterozygosity of 2.12% and repeat content of 42.88% (Figure 2). These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 52× coverage. Hi-C sequencing produced 112.83 Gb from 747.19 million reads, which were used to scaffold the assembly. Table 1 summarises the specimen and sequencing details.

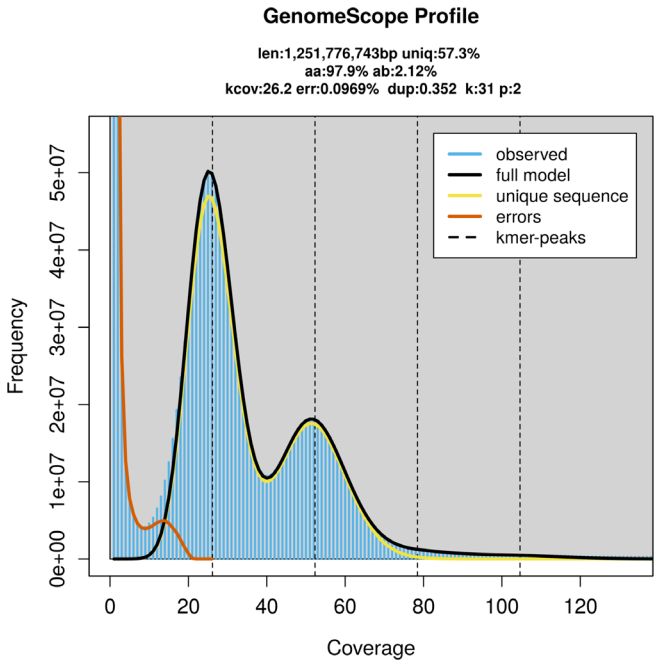

Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

Table 1. Specimen and sequencing data for BioProject PRJEB80831.

Platform	PacBio HiFi	Hi-C
ToLID	xbAmeMedi1	xbAmeMedi1
Specimen ID	SAN20002327	SAN20002327
BioSample (source individual)	SAMEA115336917	SAMEA115336917
BioSample (tissue)	SAMEA115336924	SAMEA115336924
Tissue	Whole organism	Whole organism
Instrument	Revio	Illumina NovaSeq X
Run accessions	ERR13762720	ERR13766898
Read count total	6.66 million	747.19 million
Base count total	67.68 Gb	112.83 Gb

Assembly statistics

The genome was assembled into two haplotypes using Hi-C phasing. Haplotype 1 was curated to chromosome level, while haplotype 2 was assembled to scaffold level. The final assembly has a total length of 1 299.87 Mb in 267 scaffolds, with 365 gaps, and a scaffold N50 of 67.44 Mb (Table 2).

Most of the assembly sequence (99.42%) was assigned to 19 chromosomal-level scaffolds. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 3; Table 3). During curation, we observed haplotypic inversions in the following regions: chromosome 3 (4.6–53.7 Mbp) and 5 (21.3–51.2 Mbp).

Table 2. Genome assembly statistics.

Assembly name	xbAmeMedi1.hap1.1	xbAmeMedi1.hap2.1
Assembly accession	GCA_964300355.1	GCA_964300375.1
Assembly level	chromosome	scaffold
Span (Mb)	1 299.87	1 284.99
Number of chromosomes	19	Scaffold-level
Number of contigs	632	507
Contig N50	5.6 Mb	5.99 Mb
Number of scaffolds	267	174
Scaffold N50	67.44 Mb	66.77 Mb
Longest scaffold length (Mb)	105.39	-
Organelles	Mitochondrion: 47.2 kb	-

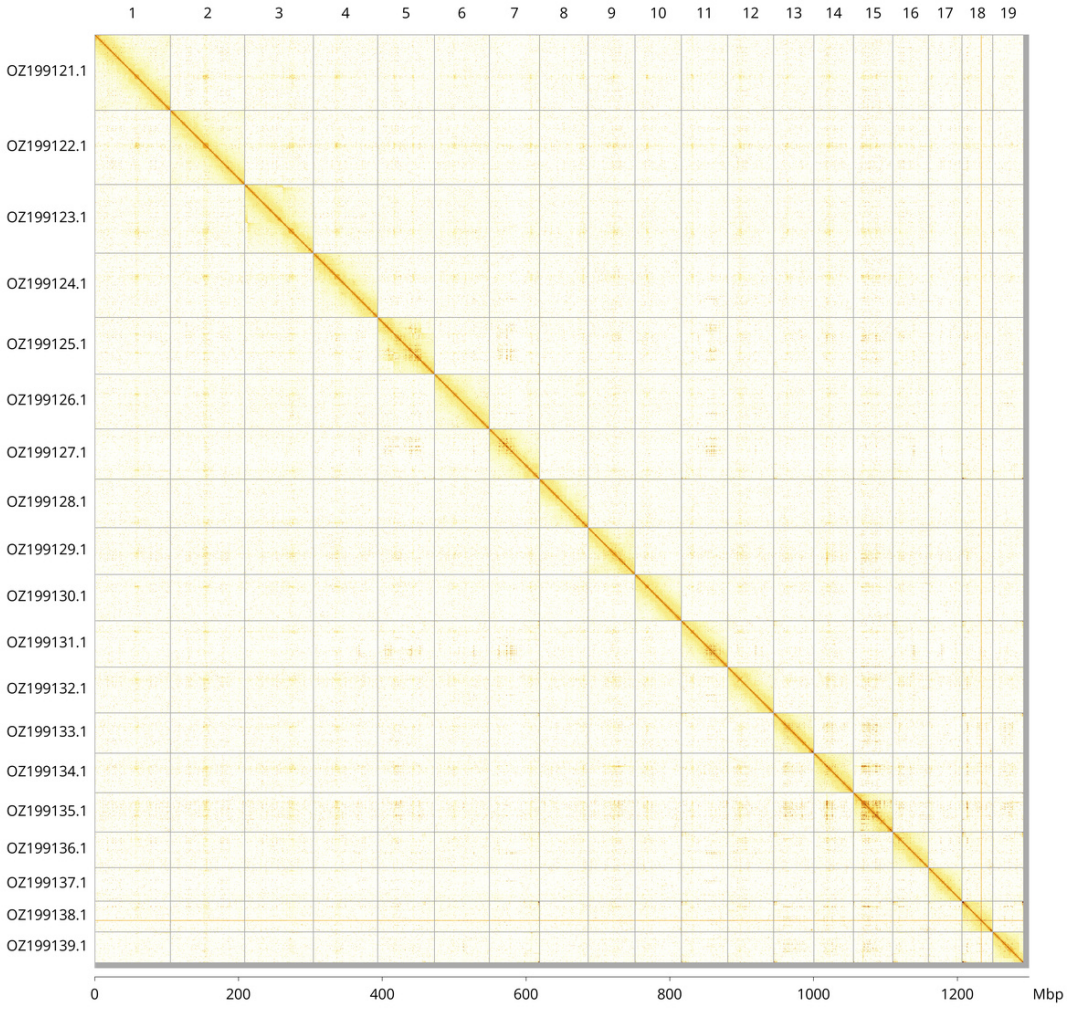

Figure 3. Hi-C contact map of the *Americardia media* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the haplotype 1 genome assembly of *Americardia media* xbAmeMedi1.

INSDC accession	Molecule	Length (Mb)	GC%
OZ199121.1	1	105.39	36
OZ199122.1	2	103.60	36
OZ199123.1	3	95.37	36
OZ199124.1	4	89.52	36.50
OZ199125.1	5	78.89	36.50
OZ199126.1	6	76.43	36
OZ199127.1	7	69.81	36
OZ199128.1	8	67.44	36
OZ199129.1	9	65.20	36.50
OZ199130.1	10	64.76	36.50
OZ199131.1	11	64.19	36.50
OZ199132.1	12	64.10	36
OZ199133.1	13	55.83	36.50
OZ199134.1	14	55.18	36.50
OZ199135.1	15	54.49	37
OZ199136.1	16	49.71	36.50
OZ199137.1	17	46.47	36.50
OZ199138.1	18	43.02	36.50
OZ199139.1	19	42.96	36.50

The mitochondrial genome was also assembled (length 47.2 kb, OZ199140.1). This sequence is included as a contig in the multifasta file of the genome submission and as a standalone record.

For haplotype 1, the estimated QV is 65.9, and for haplotype 2, 66.7. When the two haplotypes are combined, the assembly achieves an estimated QV of 66.3. The *k*-mer completeness is 68.22% for haplotype 1, 67.98% for haplotype 2, and 98.90% for the combined haplotypes (Figure 4).

BUSCO analysis using the mollusca_odb10 reference set (n = 5 295) identified 79.5% of the expected gene set (single = 78.5%, duplicated = 1.0%) for haplotype 1. The snail plot in Figure 5 summarises the scaffold length distribution and other assembly statistics for haplotype 1. The blob plot in Figure 6 shows the distribution of scaffolds by GC proportion and coverage for haplotype 1.

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 haplotype 1, is **6.C.Q65**, meeting the recommended reference standard.

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

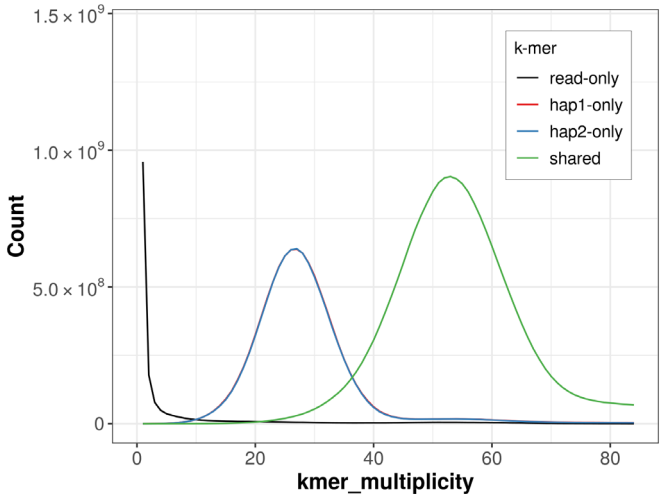

Figure 4. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers 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.

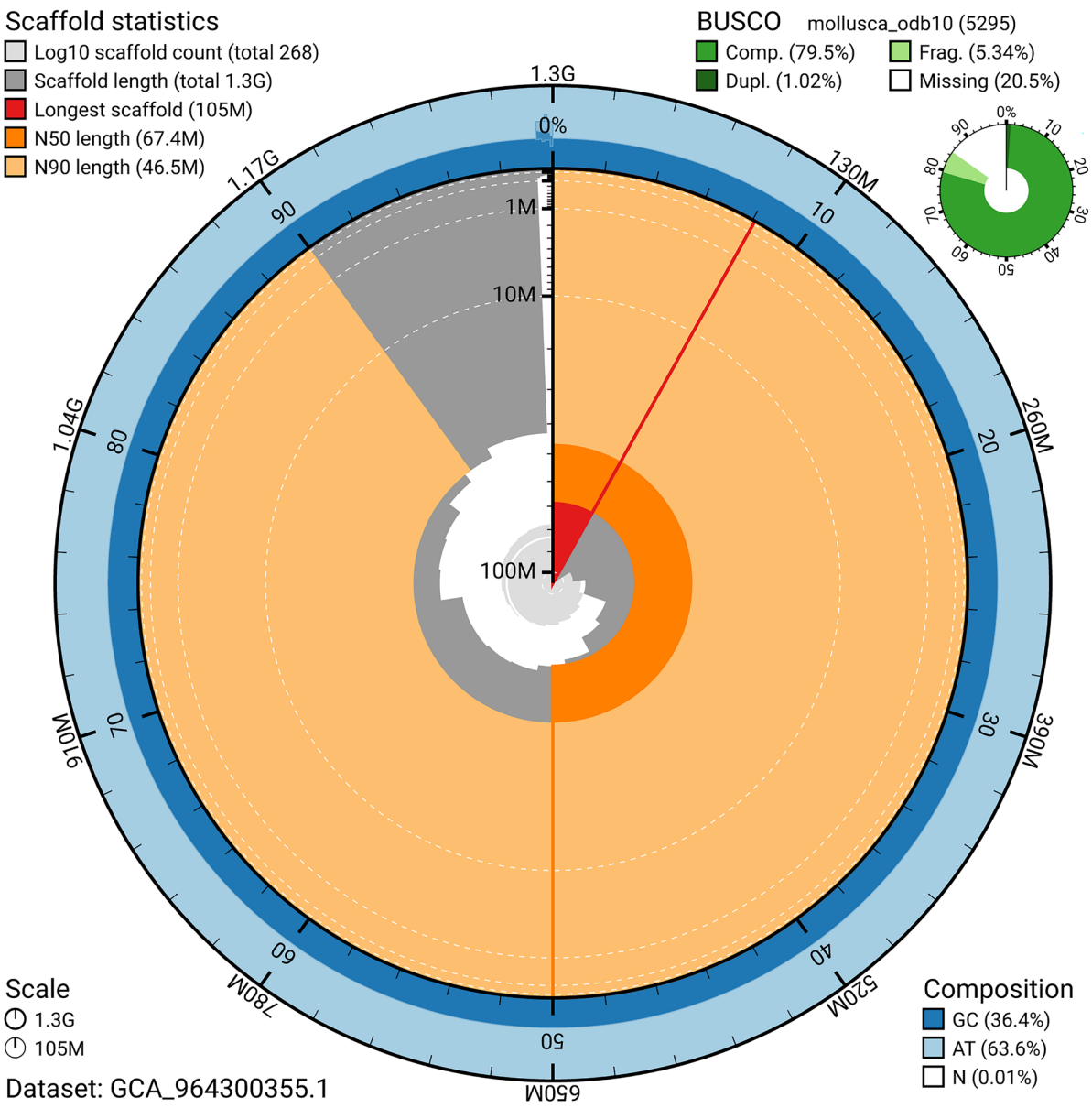

Figure 5. Assembly metrics for xbAmeMedi1.hap1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1 000 bins around the circumference. The outermost 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.

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

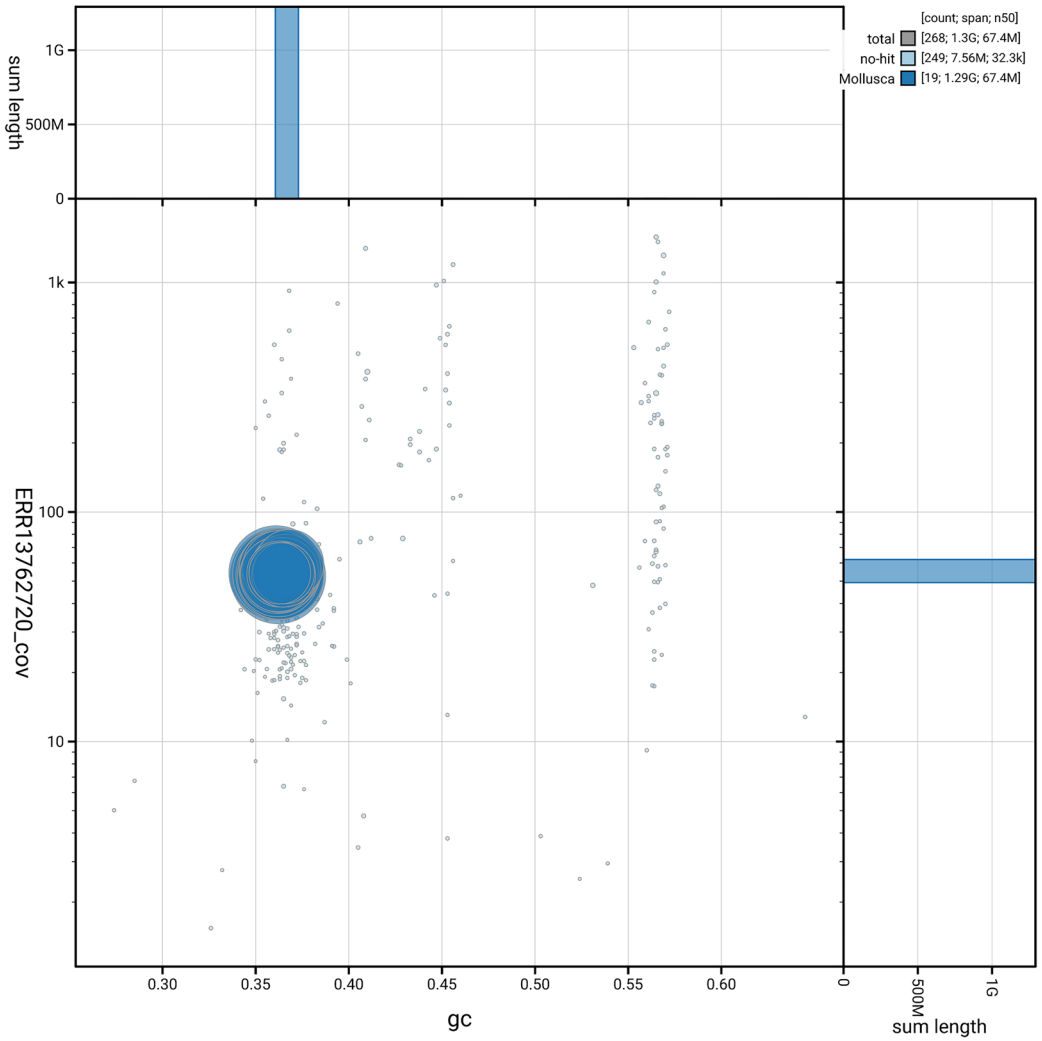

Figure 6. BlobToolKit GC-coverage plot for xbAmeMedi1.hap1.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes 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 *Americardia media* assembly.

Measure	Value	Benchmark
EBP summary (haplotype 1)	6.C.Q65	6.C.Q40
Contig N50 length	5.60 Mb	≥ 1 Mb
Scaffold N50 length	67.44 Mb	= chromosome N50
Consensus quality (QV)	Haplotype 1: 65.9; haplotype 2: 66.7; combined: 66.3	≥ 40
k-mer completeness	Haplotype 1: 68.22%; Haplotype 2: 67.98%; combined: 98.90%	≥ 95%
BUSCO	C:79.5% [S:78.5%; D:1.0%]; F:5.3%; M:15.1%; n:5 295	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	99.42%	≥ 90%

Research Limited (operating as the Wellcome Sanger Institute) and in some circumstances other Tree of Life collaborators.

Data availability

European Nucleotide Archive: Americardia media (Atlantic strawberry cockle). Accession number PRJEB80831. The genome sequence is released openly for reuse. The *Americardia media* genome sequencing initiative is part of the Sanger Institute Tree of Life Programme (PRJEB43745). All raw sequence data and the assembly have been deposited in INSDC databases. The genome will be annotated using available RNA-Seq data and presented through the Ensembl pipeline at the European Bioinformatics Institute. 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 Tree of Life Core Informatics collective

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	4.4.5	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.7.1	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
GenomeScope2.0	2.0.1	https://github.com/tbenavi1/genomescope2.0
Gfastats	1.3.6	https://github.com/vgl-hub/gfastats
GoaT CLI	0.2.5	https://github.com/genomehubs/goat-cli
Hifiasm	0.19.8-r603	https://github.com/chhylp123/hifiasm
HiGlass	1.13.4	https://github.com/higlass/higlass
MerquryFK	1.1.2	https://github.com/thegenemyers/MERQURY.FK
Minimap2	2.28-r1209	https://github.com/lh3/minimap2
MitoHiFi	3	https://github.com/marcelauliano/MitoHiFi
Oatk	1.0	https://github.com/c-zhou/oatk
MultiQC	1.14; 1.17 and 1.18	https://github.com/MultiQC/MultiQC
Nextflow	24.10.4	https://github.com/nextflow-io/nextflow
PretextSnapshot	-	https://github.com/sanger-tol/PretextSnapshot
PretextView	0.2.5	https://github.com/sanger-tol/PretextView
samtools	1.21	https://github.com/samtools/samtools
sanger-tol/ascc	0.1.0	https://github.com/sanger-tol/ascc

Software	Version	Source
sanger-tol/blobtoolkit	v0.7.1	https://github.com/sanger-tol/blobtoolkit
sanger-tol/curationpretext	1.4.2	https://github.com/sanger-tol/curationpretext
Seqtk	1.3	https://github.com/lh3/seqtk
Singularity	3.9.0	https://github.com/sylabs/singularity
TreeVal	1.4.0	https://github.com/sanger-tol/treeval
YaHS	1.2a.2	https://github.com/c-zhou/yahs

References

Allio R, Schomaker-Bastos A, Romiguier J, *et al.*: **MitoFinder: efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics.** *Mol Ecol Resour.* 2020; **20**(4): 892–905.
PubMed Abstract | Publisher Full Text | Free Full Text

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

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

Cheng H, Jarvis ED, Fedrigo O, *et al.*: **Haplotype-resolved assembly of diploid genomes without parental data.** *Nat Biotechnol.* 2022; **40**(9): 1332–1335.
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

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

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): giaa153.
PubMed Abstract | Publisher Full Text | Free Full Text

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

Kirkendale L: **Their day in the sun: molecular phylogenetics and origin of photosymbiosis in the 'other' group of photosymbiotic marine bivalves (Cardiidae: Fraginae).** *Biol J Linn Soc.* 2009; **97**(2): 448–65.
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

Li J, Lemer S, Kirkendale L, *et al.*: **Shedding light: a phylotranscriptomic perspective illuminates the origin of photosymbiosis in marine bivalves.** *BMC Evol Biol.* 2020; **20**(1): 50.
PubMed Abstract | Publisher Full Text | Free Full Text

Li R, Li J, Lemer S, *et al.*: **The genome sequence of a heart cockle, *Fragum whitleyi* Iredale, 1929 [version 1; peer review: 2 approved with reservations].** *Wellcome Open Res.* 2024a; **9**: 130.
Publisher Full Text

Li R, Li J, Lemer S, *et al.*: **The genome sequence of a heart cockle, *Fragum fragum* (Linnaeus, 1758) [version 1; peer review: 2 approved, 3 approved with reservations].** *Wellcome Open Res.* 2024b; **9**: 129.
PubMed Abstract | Publisher Full Text | Free Full Text

Li R, Li J, Lemer S, *et al.*: **The genome sequence of the heart cockle, *Fragum sueziense* (Issel, 1869) [version 1; peer review: 3 approved, 1 approved with reservations].** *Wellcome Open Res.* 2024c; **9**: 366.
PubMed Abstract | Publisher Full Text | Free 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

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

O'Leary NA, Cox E, Holmes JB, *et al.*: **Exploring and retrieving sequence and metadata for species across the Tree of Life with NCBI Datasets.** *Sci Data.* 2024; **11**(1): 732.
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.*: **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

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**

ter Poorten JJ: **A taxonomic iconography of living cardiidae**. ConchBooks, 2024.
Reference Source

Uliano-Silva M, Ferreira JGRN, Krasheninnikova 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, Brown M, Blaxter M, *et al.*: **Oatk: a de novo assembly tool for complex plant organelle genomes.** *Genome Biol.* 2025; **26**(1): 235.
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 1

Reviewer Report 31 December 2025

https://doi.org/10.21956/wellcomeopenres.27651.r139139

© **2025 Rogers R.** 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.

✔ **Rebekah L. Rogers**
UNC Charlotte, Charlotte, North Carolina, USA

This chromosomal assembly for the Atlantic Strawberry Cockle is an impressive genome that overcomes some technical challenges to achieve pseudo-chromosomal assembly of independent haplotypes for a mollusc. Genome sequencing in mollusca has often produced fragmented genomes and this genome note will be a step forward for the field.

All assembly methods and QC appear correct and straightforward to assess. The data are publicly available. This manuscript will be a fine resource release for the field of mollusc genomics.

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.

Reviewer Report 23 December 2025

https://doi.org/10.21956/wellcomeopenres.27651.r139138

© 2025 McElroy K. 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.

Kyle E McElroy
Iowa State University, Ames, USA

Summary
The authors report details on the assembly of a chromosome-level genome assembly for a species of bivalve, *Americardia media*, the Atlantic Strawberry Cockle. This species belongs to subfamily within Cardiidae that is notable for including lineages with and without photosymbiotic relationships. With genomes for multiple photosymbiotic species already available, the authors note that this non-photosymbiotic species would serve as an important contrast for comparative genomic studies. PacBio HiFi reads were generated for the initial contig assembly, which was then scaffolded into 19 chromosomal pseudomolecules with Hi-C. The genome assembly for *Americardia media* appears to be of high quality and completeness, as indicated by its contig N50, chromosome resolution, and its BUSCO and QV scores, which are all meet the Earth Biogenome Project standards. The genome was also phased into two haplotypes. The methods and results are clear and sufficiently detailed. I only have a few minor comments.

Minor comments
In the Background, I would like to see some description of the Earth Biogenome Project goals and how this resource fits in.

Given the Earth Biogenome Project benchmarks of a BUSCO completion score of >90%, the authors might want to report the metazoan db results in Table 4, which are much higher than the molluscan db results that tend to be low for even high quality assemblies.

Regarding the Hi-C methods, does the kit rely on a restriction enzyme with site-specificity or was it non-specific like DNAse I? Otherwise, I appreciate the amount of detail in the Hi-C preparations.

Out of general curiosity, was RNA-seq not generated? This line: “The genome will be annotated using available RNA-Seq data …” implies that some already exists, is that the case?

There are two mentions of mitochondrial genome assembly (from the same paragraph) that seem to be at odds with one another. Also, the mitochondrial contig should be in Table 3.

- “The mitochondrial genome was assembled using MitoHiFi (Uliano-Silva *et al.*, 2023), which …”
- “The mitochondrial genome was assembled using OATK (Zhou *et al.*, 2025).”

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 biology, genomics, genome evolution, bivalves, gene-family evolution, opsins, asexual reproduction, transposable elements

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.