scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of the striped venus clam *Chamelea gallina*

Enrico Bortoletto[1]✉, Umberto Rosani[1], Chiara Profico[2], Federica Di Giacinto[2] & Paola Venier[1]✉

The striped venus clam *Chamelea gallina* is a benthic bivalve relevant for the blue economy of the Adriatic Sea and the European fishery sector. However, the limited genomic resources available for this bivalve hinder the understanding of its biology, adaptability to climate change, and responses to stress factors. Here, we report the first chromosome-level genome assembly of *C. gallina*. Leveraging a combination of PacBio HiFi, and Illumina sequencing, complemented by Hi-C scaffolding, we generated a 1.81 Gb genome assembly with a scaffold N50 of 100.3 Mb and BUSCO completeness of 95.6%. The final assembly comprises 19 scaffolds matching the species' karyotype, demonstrating high contiguity and structural accuracy. Genome annotation identified 58,203 protein-coding genes, and repetitive regions for 54.40% of the genome. We also assembled the complete mitochondrial genome and confirmed the species identity through phylogenetic analysis. This genomic resource sets the stage for comparative analyses within the genus, the use of molecular markers for product traceability, and for the sustainable management of *C. gallina* in the context of global changes.

Background & Summary

Chamelea gallina (Linnaeus, 1758) is a *Veneridae* bivalve mollusc of about 2.5–3.5 cm in length, which thrives as a filter feeder at 5–20 m depth[1]. Morphologically, it has two symmetric valves with fine concentric ridges and typical radial stripes (Fig. 1a)[2]. *C. gallina* lives burrowing in sandy substrates (Fig. 1b)[3]. It has been reported along the Eastern Atlantic coasts, from Norway and the British Isles to the Iberian Peninsula, Morocco, and the island territories of Madeira and the Canaries (Fig. 1c). It is also a key species for the coastal blue economy of the Mediterranean and Black Seas, particularly in the Adriatic Sea[4–8].

Filtering phytoplankton and suspended matter, these clams clean the surrounding water and contribute to the element cycling, their burrowing stimulates local biodiversity, and, in turn, they are food for marine organisms like crabs and fishes. Being susceptible to environmental changes, *C. gallina* can also be regarded as an indicator species. In addition to its ecosystem services, this clam benefits the fishery sector and represents a culinary delicacy[7,9–11]. In 2023, Europe harvested 18,592 tonnes of *C. gallina*, generating around 100 million euros, with Italy being the main producer with 16,753 tonnes[12].

Several mortality events caused by unforeseen environmental changes (anoxia, river runoff, storm surges) or other factors (pollution, bacterial infections) have significantly contributed to the decline of clam beds in the Adriatic Sea[13,14]. Notably, *C. gallina* has been identified as a potential bioindicator of microplastic contamination in coastal environments, particularly in the Black Sea, where it was shown to accumulate microplastics potentially influencing its physiological status and human consumption[15]. *C. gallina* populations of the Adriatic Sea exhibited varied resilience to heat stress depending on local environmental conditions and productivity levels. In detail, clams from high-productivity areas showed enhanced antioxidant and immune responses, suggesting potential adaptation mechanisms, while the clams from low-productivity areas exhibited metabolic stress[16].

Despite the rapid advancement of genome assembly in mollusk species, facilitated by the widespread adoption of new sequencing technologies, particularly HiFi sequencing, the molecular resources available for *C. gallina* are scarce[17,18]: one transcriptome assembly published in 2012[13] and merely 16 RNA-seq samples recorded in the NCBI Sequence Read Archive (SRA, accessed 6 December 2024) in association with two different

[1]Department of Biology, University of Padova, Padova, Italy. [2]Water Biology Unit, Istituto Zooprofilattico Sperimentale dell'Abruzzo e del Molise "G. Caporale" (IZSAM), Teramo, Italy. ✉e-mail: enrico.bortoletto@unipd.it; paola.venier@unipd.it

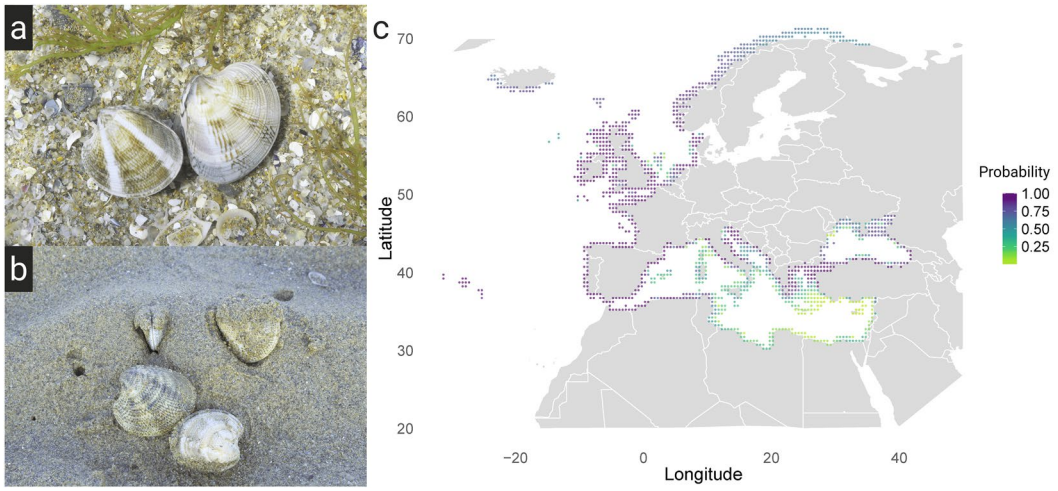

Fig. 1 Chamelea gallina traits and distribution. (**a**) Striped shells. Specimen location: Caorle, Italy. Image credit: Dr. Holger Krisp. (**b**) Burrowing habit. Specimen location: Cavallino Treporti, Italy. Image credit: Dr. Holger Krisp. (**c**) The distribution of C. gallina along the European and Mediterranean coasts. The colour gradient represents the relative probability of occurrence, from dark yellow (extremely rare) to purple (common). Distribution data sourced from AquaMaps[88].

Sample Name	Reads (Million)	Bases (Mb)	Read Layout	>Q20 (%)	>Q30 (%)	GC content (%)
HiFi[19]	7.28	135,063	SINGLE	100	98.81	36
WGS[20]	200.1	60,014	PE150	98.07	92.29	35
HI-C[21]	600.7	180,212	PE150	99.08	95.52	36

Table 1. Summary statistics of DNA sequencing samples. For each sequenced sample, the sample name, total number of reads produced, total number of bases, read layout, percentage of reads with an average quality value higher than 20 and 30, and percentage of GC content are reported.

studies[13,14]. Actually, the absence of genomic data is a critical gap in the knowledge of this species[14]. In this study, we generated Illumina short-reads, PacBio HiFi long-reads, and we used Hi-C scaffolding to construct the first chromosome-level genome assembly of *C. gallina*. This high-quality genomic resource provides a comprehensive framework for transcriptomic, epigenetic, and comparative genomics studies that could significantly expand the current knowledge of *C. gallina* biology and support effective conservation strategies.

Methods

Sample preparation. Two specimens belonging to the same population of *C. gallina* were collected from the Adriatic Sea (41°57′27.0″N, 15°04′28.1″E) on January 26th, 2024. The foot was dissected from one individual at the Istituto Zooprofilattico Sperimentale dell'Abruzzo e del Molise 'Giuseppe Caporale' (Teramo, Italy), instantly frozen in liquid nitrogen and sent to the BGI Group (Hong-Kong, China) for nucleic acid extraction and sequencing. Five tissues, including mantle, gill, hepatopancreas, hemolymph, and gonad, were dissected from a second individual collected from the same batch of animals, and immediately frozen in liquid nitrogen for RNA extraction.

DNA extraction and sequencing. To extract High Molecular Weight DNA (HMW DNA), 40 mg of tissue was snap-frozen and quickly ground into powder. The powder was dissolved in 5 mL of preheated Animal Lysis Buffer, containing 100 μL of Proteinase K solution (20 mg/mL) and 30 μL of RNase A (10 mg/mL). The reaction mixture was incubated at 55 °C for 120 min and then centrifuged at 13,000 × g for 5 min at room temperature. The supernatant was mixed with the isopropanol for the DNA precipitation, washed twice with 75% ethanol, and finally eluted in water.

For the generation of HiFi libraries HMW DNA was fragmented to an average size of 15–20 kb with the Megaruptor system (Diagenode, Seraing, Belgium). Adapters were ligated to the processed DNA to form SMRTbell™ hairpin structures. Then, linear and damaged DNA molecules were removed via enzymatic digestion to ensure high-quality library preparation. Target-sized DNA fragments were selected using BluePippin (Sage Science, Beverly, Massachusetts, USA) to optimise the read length distribution. Qualified SMRTbell™ libraries were sequenced on the PacBio Revio platform with a final output of 135,063,087,903 bases (Table 1).

A DNA library for Whole Genome Shotgun sequencing (WGS) was prepared by fragmenting 500 ng of DNA with a Covaris instrument (Covaris LLC, Woburn, Massachusetts, USA). Next, end-repair, dA tailing, magnetic bead purification, PCR amplification by KAPA HiFi HotStart DNA Polymerase (Roche, Basil, Switzerland), purification, library quality control, library circularisation were performed. The library was amplified to make

Platform	Tissue	Total Reads	Average Read Quality	Read Layout	GC content (%)
Illumina RNA-seq	Gonad[33]	94,121,308	39	PE150	41
Illumina RNA-seq	Mantle[34]	82,778,526	39	PE150	40
Illumina RNA-seq	Hepatopancreas[35]	96,384,935	39	PE150	40
Illumina RNA-seq	Gill[36]	70,618,584	39	PE150	39
Illumina RNA-seq	Hemolymph[37]	78,922,155	39	PE150	43
ONT RNA-seq	Pool[54]	471,767	9.8	860 bp	39

Table 2. Summary statistics of RNA sequencing samples. For each sequenced sample, the platform, tissue of origin, total amount of reads produced, average read quality, read layout, and percentage of GC content are reported. In the case of Nanopore sample as Read Layout we reported the average read length.

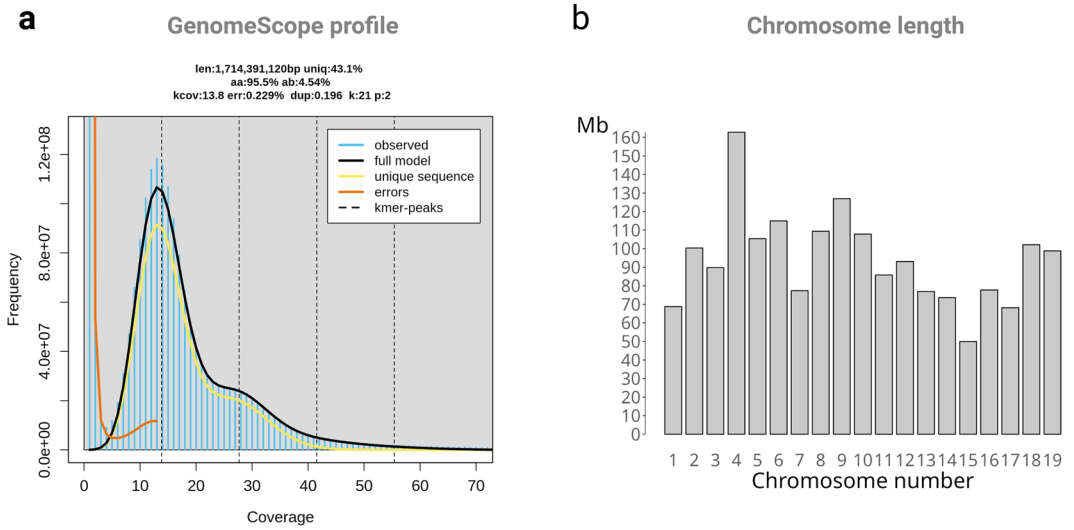

Fig. 2 Chamelea gallina genome statistics. (**a**) The GenomeScope 2.0 profile is based on k-mers (length 21) derived from short reads (Illumina whole-genome sequencing). (**b**) The histogram illustrates the dimension in Megabases (Mb) of the nineteen C. gallina chromosomes.

DNA nanoballs, and sequenced on a DNBSEQ-G400 (DNBSEQ Technology) platform, obtaining 200.1 Million reads (Table 1).

For Hi-C sequencing, a total of 40 mg of tissue was finely ground employing liquid nitrogen to ensure complete disruption of the cells. Chromatin was crosslinked with formaldehyde, then lysed, and the HindIII restriction enzyme was used to fragment the DNA at specific sites. The resulting DNA fragments were labelled with biotin to facilitate subsequent enrichment. Biotin-labelled DNA fragments were ligated to form proximity-based chimeric DNA molecules, followed by de-crosslinking with SDS and Proteinase K. The purified DNA was captured using streptavidin-coated magnetic beads. After the adapter ligation, the DNA was amplified, and the resulting libraries were subjected to quality control assessments to ensure high sequencing fidelity. Single-stranded circular DNA molecules were generated through rolling circle amplification, forming DNA nanoballs (DNBs). These DNBs were then loaded onto patterned nanoarrays and sequenced with the DNBSEQ-G400, resulting in 600.7 million reads (Table 1).

Quality assessment was performed using a custom script available in the GitHub repository associated with this publication, alongside FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Specifically, the HiFi reads[19] showed exceptional quality, with 100% of bases above Q20 and 98.81% above Q30. Similarly, the WGS[20] and Hi-C[21] datasets demonstrated high base quality, with 99.08% and 98.07% of bases above Q20, and 95.52% and 92.29% above Q30, respectively. GC content was consistent across datasets, ranging between 35% and 36% (Table 1).

RNA extraction and transcriptome sequencing. Total RNA was extracted from the five different tissues with Trizol reagent (Thermo Fisher Scientific, Waltham, Massachusetts, USA), according to the manufacturer's protocol, and the RNA quality was assessed on an Agilent TapeStation Instrument (Agilent Technologies, Palo Alto, California, USA). For short-read libraries, mRNA was enriched from total RNA following the Illumina Stranded mRNA Prep Kit instructions (Illumina Inc., San Diego, California, USA). Briefly, mRNA was captured and purified using oligo(dT) magnetic beads, selectively isolating polyadenylated RNAs. The purified mRNA was then fragmented and reverse-transcribed into first-strand cDNA using random primers. Strand specificity was achieved during second-strand synthesis by incorporating dUTP instead of dTTP. After the adapter ligation, the libraries were purified, size-selected, and amplified via PCR before sequencing on an Illumina NovaSeq X Plus

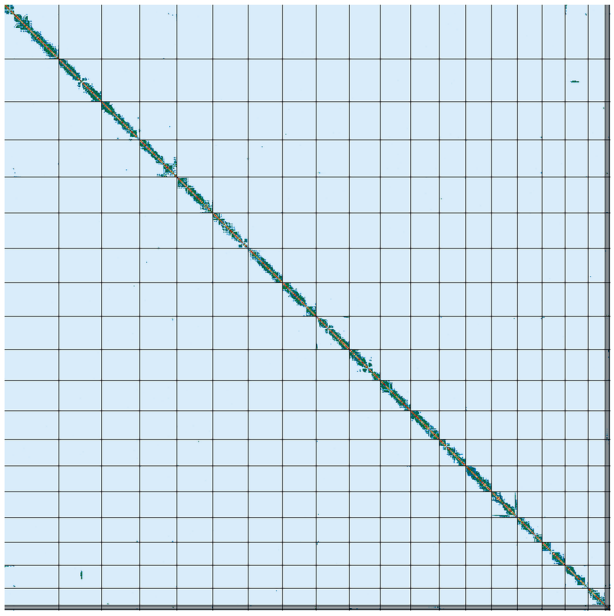

Fig. 3 Chamelea gallina contact map. Hi-C contact map for the reference genome assembly illustrates the spatial organisation of genomic regions in a linear format. Each cell in the map represents sequencing data that validates the interaction between two specific regions. The scaffolds are ordered by length and distinctly separated by black lines.

	Reference	Alternate
Total assembly length (bp)	1,807,429,086	1,715,643,596
PacBio HiFi read coverage	41x	
Number of scaffolds	37	12,454
Average scaffold length (bp)	48,849,434	137,758
Scaffold N50 (bp)	100,346,442	328,337
Largest scaffold (bp)	162,747,061	2,225,400
Contigs Number	220	12,525
Average contig length (bp)	8,215,490	136,977
Contig N50 (bp)	24,259,675	327,158
Contig L50 (number)	17	1,578
Largest contig (bp)	109,394,785	2,225,400
Base composition (%)	A: 32.4%, C: 17.6%, G: 17.6%, T: 32.4%	A: 32,5%, C: 17.5%, G: 17.5%, T: 32.5%
GC content (%)	35.14	35.09
Soft-masked bases (bp)	983,277,824	0

Table 3. Statistics of the reference and alternate genome assemblies. The Total assembly length, PacBio HiFi read coverage, Number of scaffolds, Average scaffold length, Scaffold N50, Largest scaffold, Contigs Number, Average contig length, Contig N50, Contig L50, Largest contig, Base composition, GC content % and the number of soft-masked bases are reported.

platform (Illumina Inc., San Diego, California, USA) using a paired-end 150 bp (PE150) read configuration and generating an average of 84,565,101 high-quality reads per sample (Table 2).

For the long-read RNA sequencing, we used 800 ng of total RNA to produce a single library (SQK-RNA001, Oxford Nanopore Technologies, Oxford, UK), according to the manufacturer's instructions. The poly(T) adapter was ligated to the mRNA using the T4 DNA ligase in the Quick Ligase reaction buffer (New England Biolabs, Ipswich, Massachusetts, USA) for 15 min at room temperature. First-strand cDNA was synthesised by SuperScript III Reverse Transcriptase (Thermo Fisher Scientific, Waltham, Massachusetts, USA) using the oligo(dT) adapter, and RNA–cDNA hybrids were purified using Agencourt RNAClean XP magnetic beads (Beckman Coulter, Brea, California, USA). The sequencing adapter was ligated to the mRNA using the T4 DNA ligase in the Quick Ligase reaction buffer (New England Biolabs, Ipswich, Massachusetts, USA) for 15 min at room temperature, followed by a second purification step using Agencourt beads. Nanopore sequencing of the libraries was carried out on a MinION device (Oxford Nanopore Technologies, Oxford, UK). The final sequencing output resulted in 471,767 reads (Table 2).

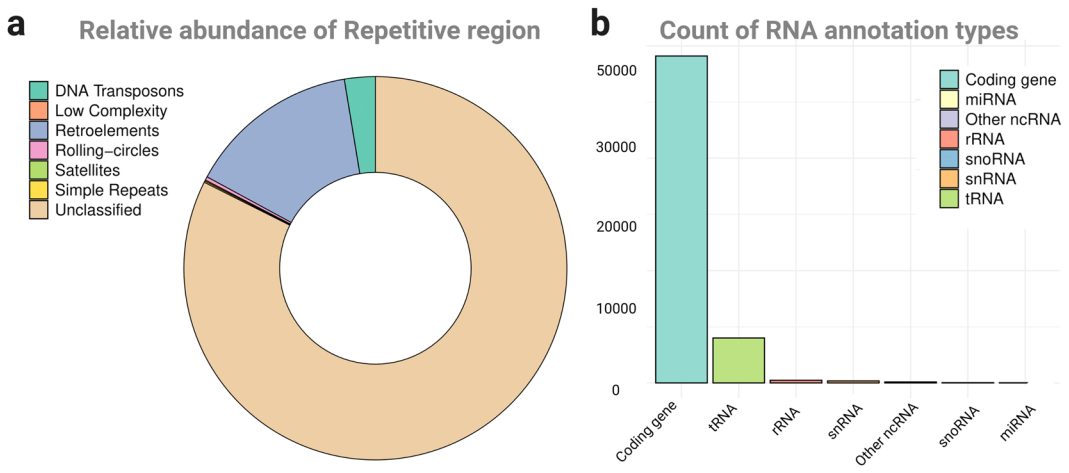

Fig. 4 Summary of the genome annotation features. (**a**) The donut chart illustrates the relative abundance of major repeat families in the Chamelea gallina genome, as identified by RepeatMasker. Different colours represent distinct repeat classes, including Retrotransposons, DNA transposons, low complexity regions, Retroelements, rolling-circle elements, Satellites, Simple, and Unclassified repeats. (**b**) The bar plot presents the distribution of annotated RNA molecules, showing the number of sequences assigned to each RNA category, including mRNA, tRNA, rRNA, snoRNA, snRNA, miRNA, and other non-coding RNAs. Different colours highlight distinct RNA types.

Assembly	Scientific Name
GCF_000002075.1[61]	*Aplysia californica*
GCF_000327385.1[62]	*Lottia gigantea*
GCF_001194135.2[63]	*Octopus bimaculoides*
GCF_002022765.2[64]	*Crassostrea virginica*
GCF_002113885.1[65]	*Mizuhopecten yessoensis*
GCF_003073045.1[66]	*Pomacea canaliculata*
GCF_003918875.1[67]	*Haliotis rubra*
GCF_006345805.1[68]	*Octopus sinensis*
GCF_016097555.1[69]	*Gigantopelta aegis*
GCF_020536995.1[70]	*Dreissena polymorpha*
GCF_021730395.1[71]	*Mercenaria mercenaria*
GCF_021869535.1[72]	*Mytilus californianus*
GCF_023055435.1[73]	*Haliotis rufescens*
GCF_025612915.1[74]	*Magallana angulata*
GCF_026571515.1[75]	*Ruditapes philippinarum*
GCF_026914265.1[76]	*Mya arenaria*
GCF_028476545.1[77]	*Physella acuta*
GCF_031769215.1[78]	*Ylistrum balloti*
GCF_032062105.1[79]	*Saccostrea cuccullata*
GCF_033153115.1[80]	*Saccostrea echinata*
GCF_036588685.1[81]	*Mytilus trossulus*
GCF_902652985.1[82]	*Pecten maximus*
GCF_902806645.1[83]	*Magallana gigas*
GCF_932274485.2[84]	*Patella vulgata*
GCF_947242115.1[85]	*Biomphalaria glabrata*
GCF_947568905.1[86]	*Ostrea edulis*
GCF_000001405.40[87]	*Homo sapiens*

Table 4. List of genome assemblies used to build the proteome database, subsequently integrated into the BRAKER3 annotation pipeline. For each genome assembly and scientific name are provided.

Genome assembly. We used *Meryl* v1.3[22] to generate k-mer counts (k = 21) from Illumina WGS reads[20]. The resulting k-mer database was then analyzed by *GenomeScope2.0*[23] to estimate sequencing errors, genome size, heterozygosity, and repeat content. As a result, the GenomeScope profile indicated a genome size of 1.71 Gb and 4.54% heterozygosity (Fig. 2a).

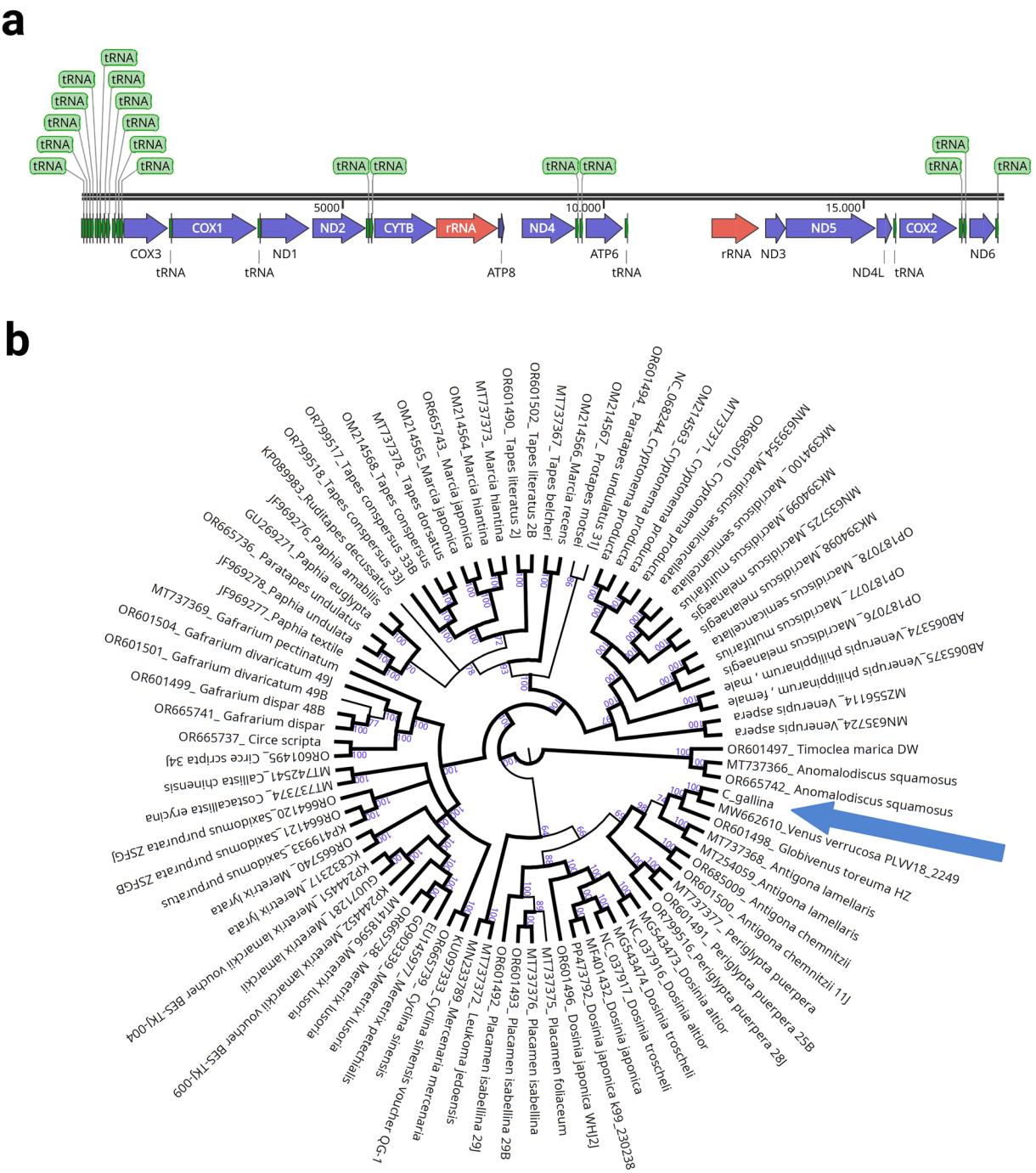

Fig. 5 Mitochondrial genome assembly. (**a**) Genomic features associated with specific positions of the Chamelea gallina mitochondrial genome. The tRNA, rRNA, and coding genes are in green, red, and blue, respectively. (**b**) Phylogenetic tree including 84 mitochondrial sequences from the Veneridae.

The genome assembly was performed using *hifiasm* v0.19.8-r603[24], with the following parameters: -s 0.35 --primary. The final results consisted of two different pseudohaplotypes, a primary assembly and an alternate assembly. The resulting primary assembly showed a size of 1.95 Gb, split into 397 different contigs with an N50 of 47.78 Mb, whereas the alternate assembly displayed a total length of 2.93 Gb, split into 31,362 contigs.

To identify and remove duplicated sequences and overlapping contigs, we applied *purge_dups* v1.2.5[25]. The forward and reverse Hi-C datasets[21] were separately mapped on both assemblies using *bwa* v0.7.18-r1243[26] with default parameters. Next, we proceeded with the scaffolding of the primary assembly using the following pipeline: the mapping BAM files were filtered and combined with the custom perl script provided by the Arima pipeline (https://github.com/ArimaGenomics/mapping_pipeline): all the reads were assigned to a single new read-group and duplicated reads were marked using picard v3.3.0 (https://github.com/broadinstitute/picard). The preprocessed BAM file was then used for the scaffolding with *SALSA* v2.3[27]. Gaps remaining after

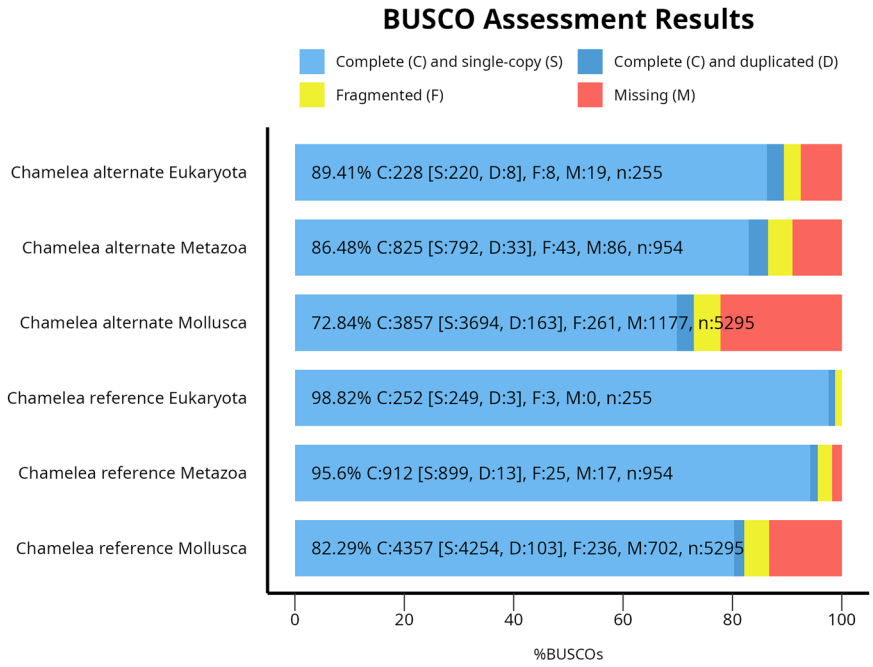

Fig. 6 BUSCO results for the assemblies. The BUSCO profiles depict the percentage of completeness, number of complete genes (C), single copy (S), duplicated (D), fragmented (F), and missing genes (M), for both the reference and alternate assemblies analysed against three different BUSCO databases (eukaryota_odb10, metazoa_odb10, and mollusca_odb10). In addition, the total number of genes included in each BUSCO database is reported (n).

scaffolding were closed using PacBio HiFi reads[19] and *YAGCloser* v1 (https://github.com/merlyescalona/yag-closer). To visually evaluate the quality of the scaffolding, the Hi-C data[21] were mapped on the scaffolded assembly and, finally, a Hi-C contact map was generated using *pairtools* v1.1.0[28] and *PretextView* v0.0.4 (https://github.com/sanger-tol/PretextView/tree/master). Previous karyotyping analysis clearly showed that *C. gallina* has 19 chromosomes[29,30]. Hi-C-based scaffolding enabled the reconstruction of 19 long pseudo-chromosomes, with lengths consistent with a chromosome-level assembly (Fig. 2b). The resulting Hi-C contact map indicates a highly contiguous assembly, free from major assembly artifacts, such as translocations or inversions (Fig. 3). Eventually, the scaffolded primary assembly (hereafter referred as reference) showed a size of 1.81 Gb and was composed of 37 scaffolds, with an N50 of 100.3 Mb (Table 3). The alternate assembly was characterised by a total length of 1.72 Gb, split into 12,454 scaffolds.

Genome annotation. To identify and mask repetitive regions, a database of repeats was built by exploiting the *RepeatModeler* v2.0.5[31] algorithm with the "-LTRStruct" option. The newly constructed database was used to softmask the genome using *RepeatMasker* v4.1.5[32] with the following parameters: "-pa 150 -q -gff --norna -no_is -e ncbi -xsmall -nolow". The genome includes a total of 983.278 Mb of repetitive regions (54.40% of the genome), composed of 14.5% of retroelements, 2.6% of DNA transposons, 0.3% of rolling-circles, 0.1% of Simple Repeats, and 0.1% of Satellite Repeats (Fig. 4a), with the great majority of repeats remaining unclassified (82.4%).

The gene models were predicted by homology modelling in combination with RNA-seq data[33–37]. In detail, the RNA-seq datasets produced from the five different tissues (mantle, gill, hepatopancreas, hemolymph, and gonad) were mapped on the final genome assembly with *hisat2* v2.2.1[38]. Next, the resulting BAM files together with a selection of proteomes (Table 4) were used to generate a first draft of gene models using *BRAKER3* v3.0.8[39] ("--skip_fixing_broken_genes").

To refine the gene models, we used the genome-guided assembly mode of *Trinity* v2.15.2[40], thus generating a transcriptome from the five RNA-seq datasets[33–37]. Then, a transcriptome database was generated with *PASA* v2.5.3[41]. *Ab-initio* gene predictions and transcript alignments were combined into weighted consensus gene structures using *EVidenceModeler* v1.1.1[42]. The non-coding genes were predicted using infernal v1.1.5[43] *cmscan* based on the Rfam library v15.0[44]. Overall, 58,203 coding genes were annotated, 8,023 tRNA, 483 rRNA, 357 snRNAs, 174 other ncRNAs, 53 snoRNAs, and 30 miRNAs (Fig. 4b).

Among 58,203 coding mRNAs, 3,356 were supported by at least 2 longreads, further validating the gene models. Functional annotation using *InterProScan* v5.57–90.0[45] identified 28,176, 30,734, and 34,385 matches with PFAM, PANTHER, and SUPERFAMILY databases, respectively. Notably, 10,329 genes lacked any functional annotation. Blastp results show that 26,243 show a hit against swissprot database.

Mitochondrial genome assembly. The mitochondrial (mtDNA) genome from the raw PacBio HiFi reads[19] was assembled following the *MitoHiFi* v3.2.2 pipeline[46], which is a reference-guided method. First, the

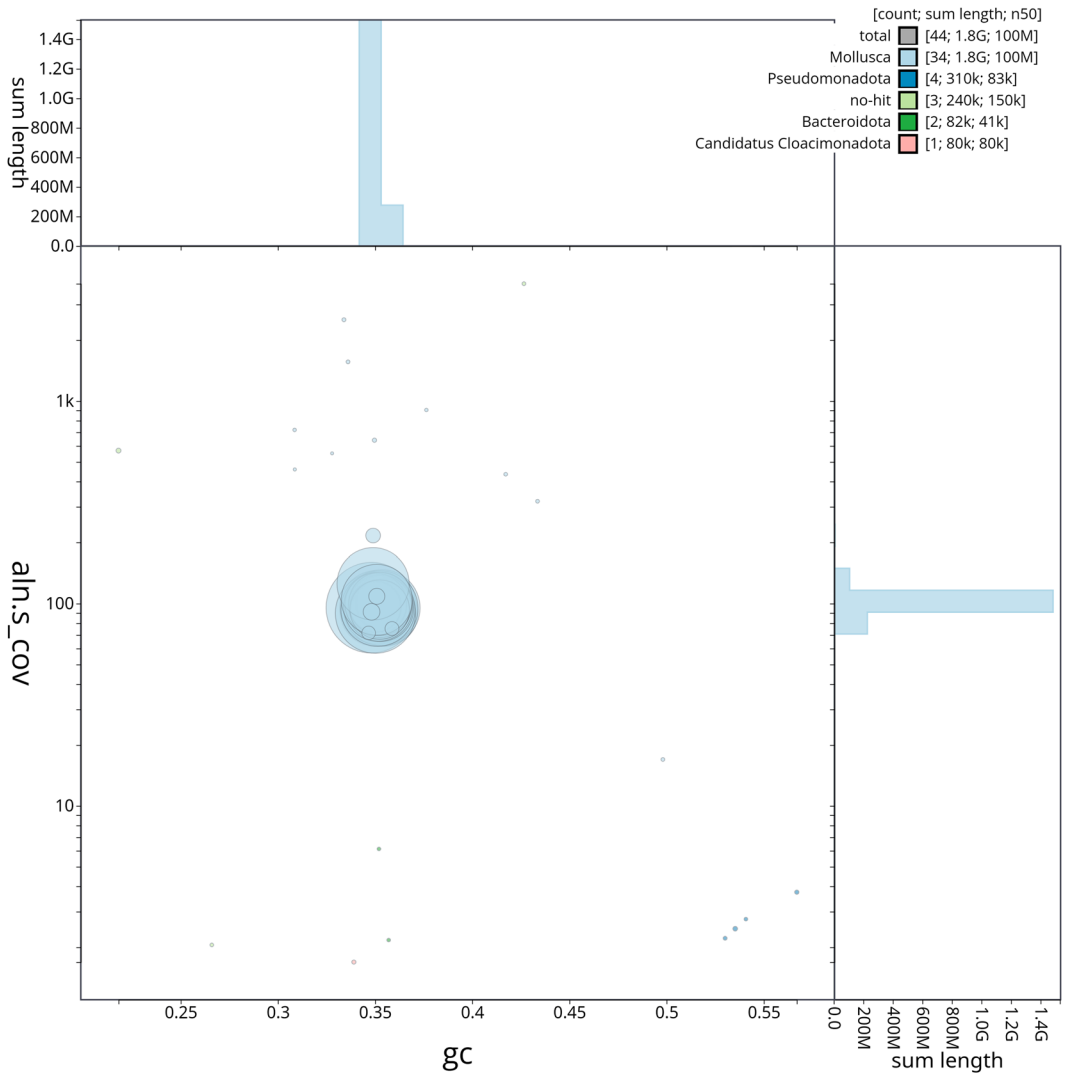

Fig. 7 Blob plot of base coverage against GC proportion for sequences of the Chamelea gallina assembly. Sequences are coloured by phylum (see colour legend). Circles are sized in proportion to sequence length. Histograms show the distribution of the sequence length sum along each axis.

software *MitoFinder* v1.4.0[47] was used for scanning all the mtDNA assemblies available on NCBI and downloading the highest quality reference mitogenome most closely related to the species of interest (in our case MW662591.1). PacBio HiFi reads[19] were mapped to the reference mitogenome using *minimap2* v2.19-r1057[48]. Then, all the mapping reads were assembled *de novo* with *hifiasm* v0.19.5-r587[24] into a 17,657 bp mitochondrial genome containing 37 genes, including 13 protein-coding genes, 2 rRNA genes, and 22 tRNA genes, with two copies of tRNA-Leu and tRNA-Ser. The genome size aligns with the expected mitochondrial genome size of eukaryotic organisms, as does the number of predicted genes (Fig. 5a). The sequencing coverage ranges from 194 to 396, ensuring sufficient depth for generating a high-quality mitochondrial genome. Then, we built a phylogenetic tree with 84 mtDNA from different bivalves, and the phylogenetic position of the newly assembled mtDNA confirms that our organism is *C. gallina* (Fig. 5b).

Statistical analysis. All statistical analyses, data manipulation and visualisation steps were performed in R 4.2.3[49] with tidyverse[50], data.table[51], magrittr[52] and, ggplot2[53] packages.

Data Record

All data, including the genome assembly raw data, the annotated reference genome assembly, the alternate assembly, and the transcriptomic data was deposited in ENA under the Umbrella project PRJEB88027. In detail, the raw genomic data are associated with the project number PRJEB87762 (HiFi data ERR14791186[19]; WGS reads ERR14790231[20]; Hi-C reads ERR14790351[21]), the transcriptomic data are available in the project PRJEB87709 (Illumina RNA seq ERR14789968[33], ERR14789500[34], ERR14789988[35], ERR14789423[36], ERR14789969[37]; Oxford Nanopore Technology data ERR14788683[54]), whereas reference and alternate genome are associated with the Genbank code GCA_965234305.2[55] and GCA_965234295.1[56], respectively. The genome

Platform	Sequence type	Tissue of origin	Percentage Mapped
DNB-seq	DNA	Foot	98.40
Illumina	RNA	Gonad	84.64
Illumina	RNA	Mantle	85.02
Illumina	RNA	Hepatopancreas	85.61
Illumina	RNA	Gill	83.28
Illumina	RNA	Hemolymph	86.84
ONT	RNA	Pool	87.86

Table 5. Summary statistics for the mapping rate. The platform, Sequence type, tissue of origin, and percentage of mapped reads are reported for each sequenced sample.

assembly produced in this work, the custom and curated GFF annotation, and protein coding files are available on figshare[57].

Technical Validation

The correctness of the genome assembly was assessed using *Flagger* v1.1.0[58]. In brief, HiFi reads were aligned to the reference genome with *minimap2* v2.28-r1209[48] with the following parameters "-t 24 --secondary = yes -ax map-pb --MD". Coverage data were extracted from the resulting BAM file and analyzed using Flagger Hidden Markov Model, which identifies local anomalies in read depth indicative of collapsed, duplicated, or erroneous regions. *Flagger* classified 0.05% of the genome as duplicated, and 0.2% as potentially erroneous. To further investigate these potentially misassembled regions, we examined their associated features and found that 71% were linked to unclassified repetitive elements. Notably, 49% corresponded to a repeat family predicted *de novo* using *RepeatModeler* v2.0.5 matching, in the NCBI *nt* database, only one significant hit, from *Venus verrucosa*, a bivalve closely related phylogenetically to *Chamelea gallina*. These specific regions may represent peri/centromeric satellite sequences, which are known to exhibit coverage biases[58].

The completeness of the reference and alternate assemblies was evaluated by the *BUSCO* v5.4.6 pipeline[59] with the following parameters -c 8 -m geno using three different databases, namely eukaryota_odb10, metazoa_odb10, and mollusca_odb10 (Fig. 6). In the case of the reference assembly, the *BUSCO* v5.4.6 pipeline indicated 98.8%, 95.6%, and 82.29% completeness for the eukaryota, metazoa, and mollusca databases, respectively. Lower values were obtained in the case of the alternate assembly (89.41% eukaryota, 86.48% metazoa, and 72.84% mollusca).

Overall, the *BUSCO* summary supports the high quality of the assembled genomes (Fig. 6). Moreover, the k-mer spectrum of the WGS read set[20] had been used to independently evaluate the assembly quality without relying on a high-quality reference. In brief, by comparing k-mers in the de novo assembly to those found in unassembled high-accuracy reads, we estimated the consensus quality (QV) and k-mer completeness using *merqury* v1.3[22]. Relevant as key metrics, we obtained a k-mer completeness of 65.4, and a k-mer-based QV of 46.4 for the reference assembly. Also in this case, the alternate assembly displayed lower values with k-mer completeness of 63.8, and a k-mer-based QV of 46.8. Considering both assemblies together, the estimated k-mer completeness rocketed to 99.4. We applied the blobtoolkit pipeline[60] to assess the genome contamination level, and the resulting Blob plot allowed us to remove the contaminant scaffolds (Fig. 7).

As an independent evaluation of the genome quality, we mapped the WGS data[20] produced to the reference assembly and obtained a 98.4% mapping rate. Similarly, the majority of RNA reads obtained from single tissues or pooled tissues mapped on the genome (from 83.28% to 87.86%) (Table 5).

Data availability

All data, including the genome assembly raw data, the annotated reference genome assembly, the alternate assembly, and the transcriptomic data was deposited in ENA under the Umbrella project PRJEB88027 https://www.ebi.ac.uk/ena/browser/view/PRJEB88027. The genome assembly produced and analyzed in this work, the custom and curated GFF annotation, and protein coding files are available on figshare https://figshare.com/articles/online_resource/Chamelea_gallina_genome_assembly/28789796/7.

Code availability

All the tools and software used in this study are described and cited in the Methods section. All the commands, including the applied parameters, are available in the GitHub repository: https://github.com/GitEnricoNeko/Chamelea_gallina_assembly_pipeline.

Received: 23 May 2025; Accepted: 26 January 2026;
Published online: 11 February 2026

References

1. Striped venus - Aquatic species. https://www.fao.org/fishery/en/aqspecies/2697/en.
2. Gizzi, F. *et al*. Shell properties of commercial clam Chamelea gallina are influenced by temperature and solar radiation along a wide latitudinal gradient. *Sci. Rep.* **6**, 36420, https://doi.org/10.1038/srep36420 (2016).
3. Grazioli, E. *et al*. Review of the Scientific Literature on Biology, Ecology, and Aspects Related to the Fishing Sector of the Striped Venus (Chamelea gallina) in Northern Adriatic Sea. *J. Mar. Sci. Eng.* **10**, 1328, https://doi.org/10.3390/jmse10091328 (2022).
4. Gaspar, M. B. & Monteiro, C. C. Reproductive Cycleso Of the Razor Clam Ensis Siliqua and the Clam Venus Striatula off Vilamoura, Southern Portugal. *J. Mar. Biol. Assoc. U. K.* **78**, 1247–1258, https://doi.org/10.1017/S0025315400044465 (1998).

5. Orban, E. *et al*. Nutritional and commercial quality of the striped venus clam, *Chamelea gallina*, from the Adriatic sea. *Food Chem.* **101**, 1063–1070, https://doi.org/10.1016/j.foodchem.2006.03.005 (2007).
6. Chamelea gallina, Striped venus clam: fisheries. https://www.sealifebase.se/summary/Chamelea-gallina.html.
7. Carducci, F. *et al*. Omics approaches for conservation biology research on the bivalve Chamelea gallina. *Sci. Rep.* **10**, 19177, https://doi.org/10.1038/s41598-020-75984-9 (2020).
8. Baeta, M., Solís, M. A., Ballesteros, M. & Defeo, O. Long-term trends in striped venus clam (*Chamelea gallina*) fisheries in the western Mediterranean Sea: The case of Ebro Delta (NE Spain). *Mar. Policy* **134**, 104798, https://doi.org/10.1016/j.marpol.2021.104798 (2021).
9. Bargione, G. *et al*. Chamelea gallina reproductive biology and Minimum Conservation Reference Size: implications for fishery management in the Adriatic Sea. *BMC Zool.* **6**, 32, https://doi.org/10.1186/s40850-021-00096-4 (2021).
10. Carducci, F. *et al*. The Mantle Transcriptome of Chamelea gallina (Mollusca: Bivalvia) and Shell Biomineralization. *Animals* **12**, 1196, https://doi.org/10.3390/ani12091196 (2022).
11. La risorsa | Co.Ge.Vo. Ancona. https://www.cogevoancona.it/?page_id=2054.
12. Statistics | Eurostat. https://ec.europa.eu/eurostat/databrowser/view/fish_ld_main__custom_13569048/bookmark/table?lang=en&bookmarkId=f3bbfb5d-a492-4b7a-850f-1f7815bc13f8.
13. Coppe, A. *et al*. Sequencing and Characterization of Striped Venus Transcriptome Expand Resources for Clam Fishery Genetics. *PLOS ONE* **7**, e44185, https://doi.org/10.1371/journal.pone.0044185 (2012).
14. Milan, M. *et al*. Host-microbiota interactions shed light on mortality events in the striped venus clam Chamelea gallina. *Mol. Ecol.* **28**, 4486–4499, https://doi.org/10.1111/mec.15227 (2019).
15. Gedik, K. & Gozler, A. M. Hallmarking microplastics of sediments and *Chamelea gallina* inhabiting Southwestern Black Sea: A hypothetical look at consumption risks. *Mar. Pollut. Bull.* **174**, 113252, https://doi.org/10.1016/j.marpolbul.2021.113252 (2022).
16. Iuffrida, L. *et al*. Physiological plasticity and life history traits affect *Chamelea gallina* acclimatory responses during a marine heatwave. *Environ. Res.* **263**, 120287, https://doi.org/10.1016/j.envres.2024.120287 (2024).
17. Qi, H., Cong, R., Wang, Y., Li, L. & Zhang, G. Construction and analysis of the chromosome-level haplotype-resolved genomes of two *Crassostrea* oyster congeners: *Crassostrea angulata* and *Crassostrea gigas*. *GigaScience* **12**, giad077, https://doi.org/10.1093/gigascience/giad077 (2022).
18. Espinosa, E., Bautista, R., Larrosa, R. & Plata, O. Advancements in long-read genome sequencing technologies and algorithms. *Genomics* **116**, 110842, https://doi.org/10.1016/j.ygeno.2024.110842 (2024).
19. *The European Nucleotide Archive (ENA)* http://identifiers.org/ena.embl:ERR14791186 (2025).
20. *The European Nucleotide Archive (ENA)* http://identifiers.org/ena.embl:ERR14790231 (2025).
21. *The European Nucleotide Archive (ENA)* http://identifiers.org/ena.embl:ERR14790351 (2025).
22. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245, https://doi.org/10.1186/s13059-020-02134-9 (2020).
23. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1432, https://doi.org/10.1038/s41467-020-14998-3 (2020).
24. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175, https://doi.org/10.1038/s41592-020-01056-5 (2021).
25. Guan, D. *et al*. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898, https://doi.org/10.1093/bioinformatics/btaa025 (2020).
26. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760, https://doi.org/10.1093/bioinformatics/btp324 (2009).
27. Ghurye, J., Pop, M., Koren, S., Bickhart, D. & Chin, C.-S. Scaffolding of long read assemblies using long range contact information. *BMC Genomics* **18**, 527, https://doi.org/10.1186/s12864-017-3879-z (2017).
28. Open2C. *et al*. Pairtools: From sequencing data to chromosome contacts. *PLOS Comput. Biol.* **20**, e1012164, https://doi.org/10.1371/journal.pcbi.1012164 (2024).
29. Corni, M. G. & Trentini, M. A chromosomic study of Chamelea gallina (L.) (Bivalvia, Veneridae). *Boll. Zool.* **53**, 23–24, https://doi.org/10.1080/11250008609355477 (1986).
30. García-Souto, D., Qarkaxhija, V. & Pasantes, J. J. Resolving the Taxonomic Status of Chamelea gallina and C. striatula (Veneridae, Bivalvia): A Combined Molecular Cytogenetic and Phylogenetic Approach. *BioMed Res. Int.* **2017**, 7638790, https://doi.org/10.1155/2017/7638790 (2017).
31. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci.* **117**, 9451–9457, https://doi.org/10.1073/pnas.1921046117 (2020).
32. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinforma.* Chapter 4, 4.10.1–4.10.14, https://doi.org/10.1002/0471250953.bi0410s25 (2009).
33. *The European Nucleotide Archive (ENA)* http://identifiers.org/ena.embl:ERR14789968 (2025).
34. *The European Nucleotide Archive (ENA)* http://identifiers.org/ena.embl:ERR14789500 (2025).
35. *The European Nucleotide Archive (ENA)* http://identifiers.org/ena.embl:ERR14789988 (2025).
36. *The European Nucleotide Archive (ENA)* http://identifiers.org/ena.embl:ERR14789423 (2025).
37. *The European Nucleotide Archive (ENA)* http://identifiers.org/ena.embl:ERR14789969 (2025).
38. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915, https://doi.org/10.1038/s41587-019-0201-4 (2019).
39. Gabriel, L. *et al*. BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. Genome Res. 34, 769–777, https://doi.org/10.1101/gr.278090.123.(2024).
40. Grabherr, M. G. *et al*. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat. Biotechnol.* **29**, 644–652, https://doi.org/10.1038/nbt.1883 (2011).
41. Haas, B. J. *et al*. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666, https://doi.org/10.1093/nar/gkg770 (2003).
42. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7, https://doi.org/10.1186/gb-2008-9-1-r7 (2008).
43. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinforma. Oxf. Engl.* **29**, 2933–2935, https://doi.org/10.1093/bioinformatics/btt509 (2013).
44. Ontiveros-Palacios, N. *et al*. Rfam 15: RNA families database in 2025. *Nucleic Acids Res.* **53**, D258–D267, https://doi.org/10.1093/nar/gkae1023 (2025).
45. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240, https://doi.org/10.1093/bioinformatics/btu031 (2014).
46. Uliano-Silva, M. *et al*. MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads. *BMC Bioinformatics* **24**, 288, https://doi.org/10.1186/s12859-023-05385-y (2023).
47. Allio, R. *et al*. MitoFinder: Efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. *Mol. Ecol. Resour.* **20**, 892–905, https://doi.org/10.1111/1755-0998.13160 (2020).
48. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinforma. Oxf. Engl.* **34**, 3094–3100, https://doi.org/10.1093/bioinformatics/bty191 (2018).

49. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, Vienna, Austria, 2021).
50. Wickham, H. *et al*. Welcome to the Tidyverse. *J. Open Source Softw.* **4**, 1686, https://doi.org/10.21105/joss.01686 (2019).
51. Barrett, T. *et al*. *Data.Table: Extension of 'data.Frame'*. (2024).
52. Bache, S. M. & Wickham, H. *Magrittr: A Forward-Pipe Operator for R*. (2022).
53. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*. (Springer international publishing, Cham, 2016).
54. *The European Nucleotide Archive (ENA)* http://identifiers.org/ena.embl:ERR14788683 (2025).
55. ENA https://www.ebi.ac.uk/ena/browser/view/GCA_965234305.2 (2025).
56. ENA https://www.ebi.ac.uk/ena/browser/view/GCA_965234295.1 (2025).
57. Bortoletto, E. Chamelea gallina genome assembly. Figshare. https://doi.org/10.6084/m9.figshare.28789796.v7 (2025).
58. Liao, W.-W. *et al*. A draft human pangenome reference. *Nature* **617**, 312–324, https://doi.org/10.1038/s41586-023-05896-x (2023).
59. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol. Biol. Evol.* **38**, 4647–4654, https://doi.org/10.1093/molbev/msab199 (2021).
60. Challis, R., Richards, E., Rajan, J., Cochrane, G. & Blaxter, M. BlobToolKit – Interactive Quality Assessment of Genome Assemblies. *G3 GenesGenomesGenetics* **10**, 1361–1374, https://doi.org/10.1534/g3.119.400908 (2020).
61. *NCBI GenBank* http://identifiers.org/refseq.gcf:GCF_000002075.1 (2013).
62. *NCBI GenBank* http://identifiers.org/refseq.gcf:GCF_000327385.1 (2012).
63. *NCBI GenBank* http://identifiers.org/refseq.gcf:GCF_001194135.2 (2022).
64. *NCBI GenBank* http://identifiers.org/refseq.gcf:GCF_002022765.2 (2017).
65. *NCBI GenBank* http://identifiers.org/refseq.gcf:GCF_002113885.1 (2017).
66. *NCBI GenBank* http://identifiers.org/refseq.gcf:GCF_003073045.1 (2018).
67. *NCBI GenBank* http://identifiers.org/refseq.gcf:GCF_003918875.1 (2018).
68. *NCBI GenBank* http://identifiers.org/refseq.gcf:GCF_006345805.1 (2019).
69. *NCBI GenBank* http://identifiers.org/refseq.gcf:GCF_016097555.1 (2020).
70. *NCBI GenBank* http://identifiers.org/refseq.gcf:GCF_020536995.1 (2021).
71. *NCBI GenBank* http://identifiers.org/refseq.gcf:GCF_021730395.1 (2022).
72. *NCBI GenBank* http://identifiers.org/refseq.gcf:GCF_021869535.1 (2022).
73. *NCBI GenBank* http://identifiers.org/refseq.gcf:GCF_023055435.1 (2022).
74. *NCBI GenBank* http://identifiers.org/refseq.gcf:GCF_025612915.1 (2022).
75. *NCBI GenBank* http://identifiers.org/refseq.gcf:GCF_026571515.1 (2022).
76. *NCBI GenBank* http://identifiers.org/refseq.gcf:GCF_026914265.1 (2022).
77. *NCBI GenBank* http://identifiers.org/refseq.gcf:GCF_028476545.1 (2023).
78. *NCBI GenBank* http://identifiers.org/refseq.gcf:GCF_031769215.1 (2023).
79. *NCBI GenBank* http://identifiers.org/refseq.gcf:GCF_032062105.1 (2023).
80. *NCBI GenBank* http://identifiers.org/refseq.gcf:GCF_033153115.1 (2023).
81. *NCBI GenBank* http://identifiers.org/refseq.gcf:GCF_036588685.1 (2024).
82. *NCBI GenBank* http://identifiers.org/refseq.gcf:GCF_902652985.1 (2019).
83. *NCBI GenBank* http://identifiers.org/refseq.gcf:GCF_902806645.1 (2020).
84. *NCBI GenBank* http://identifiers.org/refseq.gcf:GCF_932274485.2 (2022).
85. *NCBI GenBank* http://identifiers.org/refseq.gcf:GCF_947242115.1 (2022).
86. *NCBI GenBank* http://identifiers.org/refseq.gcf:GCF_947568905.1 (2022).
87. *NCBI GenBank* http://identifiers.org/refseq.gcf:GCF_000001405.40 (2022).
88. Kaschner, K., Ready, J. S., Agbayani, E., Rius, J. & Kesner-Reyes, K. AquaMaps: Predicted range maps for aquatic species (2008).

Acknowledgements

This project has received funding from the European Union under the European Union's Horizon Europe research and innovation programme, under the Biodiversity, Circular Economy and Environment (REA.B.3); co-funded by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 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 genome project stems from and partially received funding from the project "Valutazione dell'esposizione a microplastiche e nanoplastiche associata al consumo di vongola *Chamelea gallina*" (PLASTICVONG, Code IZS AM 07 /22 RC, CUP B45E22001850001, granted from the Italian Ministry of Health to F. Di Giacinto, Istituto Zooprofilattico Sperimentale dell'Abruzzo e del Molise "G. Caporale, Italy. We want to thank Dr. Holger Krisp for providing nice pictures of *Chamelea gallina*. We express our gratitude to the HPC infrastructure CAPRI (Calcolo ad Alte Prestazioni per la Ricerca e l'Innovazione) for providing the computational power needed to run all the bioinformatics analyses (University of Padova Strategic Research Infrastructure Grant 2017). The PostDoc fellowship to EB was granted by the Italian Ministry of University and Research (MIUR), PRIN project 202292P4R7. The publication of this work was supported by the Italian Ministry of University and Research (MIUR), PRIN project P2022JEEMT.

Author contributions

Conceptualization: E.B., U.R., P.V.; Data curation: E.B., U.R.; Sampling: C.P., F.D.G.; Formal analysis: E.B., U.R.; Funding acquisition: P.V., F.D.G.; Investigation: E.B., U.R., P.V., C.P.; Methodology: E.B., U.R.; Project administration: E.B., P.V.; Resources: P.V., F.D.G.; Software: E.B.; Supervision: P.V., U.R.; Validation: E.B.; Visualization: E.B., U.R.; Writing - original draft: E.B.; Writing - review & editing: E.B., U.R., P.V., F.D.G., C.P.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to E.B. or P.V.