scientific data

Check for updates

OPEN

DATA DESCRIPTOR

# Chromosome-level genome assembly of the bay scallop *Argopecten irradians*

Denis Grouzdev[1], Emmanuelle Pales Espinosa[1], Stephen Tettelbach[2], Sarah Farhat[1,3], Arnaud Tanguy[4], Isabelle Boutet[4], Nadège Guiglielmoni[5], Jean-François Flot[5,6], Harrison Tobi[2] & Bassem Allam[1]✉

**The bay scallop, *Argopecten irradians*, is a species of major commercial, cultural, and ecological importance. It is endemic to the eastern coast of the United States, but has also been introduced to China, where it supports a significant aquaculture industry. Here, we provide an annotated chromosome-level reference genome assembly for the bay scallop, assembled using PacBio and Hi-C data. The total genome size is 845.9 Mb, distributed over 1,503 scaffolds with a scaffold N50 of 44.3 Mb. The majority (92.9%) of the assembled genome is contained within the 16 largest scaffolds, corresponding to the 16 chromosomes confirmed by Hi-C analysis. The assembly also includes the complete mitochondrial genome. Approximately 36.2% of the genome consists of repetitive elements. The BUSCO analysis showed a completeness of 96.2%. We identified 33,772 protein-coding genes. This genome assembly will be a valuable resource for future research on evolutionary dynamics, adaptive mechanisms, and will support genome-assisted breeding, contributing to the conservation and management of this iconic species in the face of environmental and pathogenic challenges.**

## Background & Summary

Bivalves (Bivalvia), a class of mollusks, encompass a vast array of species that are integral to marine and freshwater ecosystems[1,2]. Most members of the class are filter feeders that help mitigate eutrophication and improve water quality, and many bivalve species serve as bioindicators of environmental health and represent a vital food source for humans and other animals[3–5]. In 2020, over 16 million tons of bivalves were produced from farming activities worldwide[6], representing a commercial value of nearly 30 billion US$. Among the bivalves, the family Pectinidae, commonly known as scallops, is of particular interest due to their ecological significance and economic value[7]. The Pectinidae family comprises over 300 species distributed worldwide, with members known for their distinctive fan-shaped shells[8]. This family includes the bay scallop, *Argopecten irradians*, a species that has attracted considerable attention due to its unique biological traits and commercial importance. The species displays remarkable polymorphism in shell color patterns (Fig. 1), relatively short lifespan, and exhibits unique locomotion through rapid shell clapping[9–11].

The bay scallop naturally inhabits shallow coastal waters along the eastern coast of North America, from New England to the Gulf of Mexico[12,13]. They prefer estuaries and bays with relatively high salinity, water depths of 0.3 to 0.6 m at low tide, and seagrass beds[14]. Small batches of *A. irradians* were introduced from the United States to China in the 1980s and 1990s and served to establish a very successful aquaculture production, yielding about 1 million tons annually[15–17].

The genomic study of bivalves, particularly within the Pectinidae family, has lagged behind other groups such as oysters[18] and mussels[19], leaving a substantial gap in our understanding of their genetic diversity and adaptive potential. Scallop genomic assemblies have been previously generated for *Mizuhopecten yessoensis*[20],

[1]School of Marine and Atmospheric Sciences, Stony Brook University, Stony Brook, NY, 11794-5000, USA. [2]Cornell Cooperative Extension of Suffolk County, Southold, NY, 11971, USA. [3]Institut Systématique Evolution Biodiversité (ISYEB), Muséum national d'Histoire naturelle, CNRS, Sorbonne Université, EPHE, Université des Antilles, 57 rue Cuvier, CP 50, 75005, Paris, France. [4]Station Biologique de Roscoff, CNRS/Sorbonne Université, Place Georges Teissier, 29680, Roscoff, France. [5]Evolutionary Biology and Ecology, Université libre de Bruxelles (ULB), 1050, Brussels, Belgium. [6]Interuniversity Institute of Bioinformatics in Brussels – (IB)², Brussels, Belgium. ✉e-mail: bassem.allam@stonybrook.edu

**Fig. 1** A representation of color and pattern variations in *Argopecten irradians*, commonly known as the bay scallop. Photo: S. Tettelbach.

*Chlamys farreri*[21], and *Argopecten purpuratus*[22], but chromosome-scale scallop genome assemblies available in open databases are currently limited to *Pecten maximus*[23] and *Mimachlamys varia*[24]. Recently, the draft genomes of bay scallop subspecies (*irradians* and *concentricus*) cultivated in China have been sequenced[25]. However, these genomes are scaffolded and reflect a complex introduction history due to their aquaculture origin. Prior studies also reported reduced allele diversity in bay scallop populations in China, suggesting that the limited man-made stock introductions may have yielded a bottleneck in genetic diversity among continuously cultured stocks[26]. This complexity, in addition to the draft nature of the current assemblies, may pose challenges for using these genomes in genomic and environmental research related to the species' natural habitat.

The availability of a high-quality genome assembly for *A. irradians* marks a significant advancement in bivalve genomics, promising to shed light on the complex biological processes that underpin their survival and productivity. Especially important is the fact that since 2019, the bay scallop population in New York has suffered from catastrophic and recurring summer mortality events that has devastated the commercial fishery. This mortality is associated with annual outbreaks of an undescribed apicomplexan parasite, recently dubbed Bay Scallop Marosporida (BSM)[27,28]. This study presents the first chromosome-level genome assembly of *A. irradians*, achieved using PacBio sequencing and Hi-C technology. The assembled genome measures 845.9 Mb, featuring a scaffold N50 length of 44.3 Mb. A total of 33,772 protein-coding genes were predicted within the *A. irradians* genome. This high-quality assembly, derived from specimens in their native habitat in New York, provides a crucial genomic resource for advancing genetic improvement and elucidating the functional genes and molecular mechanisms underlying the peculiar traits of the bay scallop. In contrast to the existing *A. irradians* draft genome assemblies, this newly assembled genome offers significant improvements in both resolution and completeness, resulting in a more contiguous and comprehensive assembly. The genome was generated from a scallop produced by a breeding program that uses wild broodstock to maintain a broader genetic base, thereby reducing the bottleneck effect commonly observed in aquaculture stocks.

## Methods

**Sample collection and genome sequencing.** The reference genome was generated from an adult scallop (62 mm) collected from a first-generation aquacultured stock bred by Cornell Cooperative Extension from wild broodstock harvested from Orient Harbor, New York, USA (41.137904, −72.315392). The scallop was transported to the laboratory on ice for processing, where the testis was dissected and immediately used for DNA extraction using standard phenol-chloroform-isoamyl alcohol (PCI) extraction[29]. In parallel, the adductor muscle was dissected and flash-frozen in liquid nitrogen before transfer to a −80 °C freezer for subsequent Hi-C sequencing. High-molecular-weight gDNA obtained from testis and subsequently purified was prepared for PacBio single-molecule real-time (SMRT) sequencing using the Express Template Preparation Kit 2.0 (Pacific Biosciences) according to the manufacturer's protocol. Approximately 2 μg of gDNA was sheared to create 10-kb libraries using Covaris g-TUBEs, followed by concentration using 0.45X AMPure PB beads (Pacific Biosciences). This sheared gDNA was enzymatically treated to remove single stranded overhangs and to repair nicked DNA templates. An end repair and A-tailing reaction further prepared the sample by repairing blunt ends and polyadenylating each template. SMRTbell adapters were then ligated to each template and 0.45X AMPure PB beads were used for purification to remove small fragments and excess reagents. Size selection of the purified SMRTbell libraries was performed at 6–50 kb using the BluePippin system on 0.75% agarose cassettes and S1 ladders according to the manufacturer's specifications (Sage Science (Beverly, Massachusetts, USA)). The final size-selected library was annealed to sequencing primer v4 and coupled to sequencing polymerase 1.0, then sequenced on two 8 M SMRT cells on the Sequel II system, each with a 20-hour movie. This resulted in a total of 9,919,395 reads with an average

| Assembly | *Argopecten irradians* NY |
|---|---|
| # scaffolds (>= 0 bp) | 1,509 |
| # scaffolds (>= 5000 bp) | 1,470 |
| # scaffolds (>= 10000 bp) | 1,352 |
| # scaffolds (>= 50000 bp) | 572 |
| Total length (>= 0 bp) | 845,909,515 |
| Total length (>= 5000 bp) | 845,770,363 |
| Total length (>= 10000 bp) | 844,855,149 |
| Total length (>= 50000 bp) | 825,008,202 |
| Largest contig | 75,513,885 |
| GC (%) | 35.6 |
| N50 | 44,345,813 |
| N90 | 34,531,595 |
| L50 | 8 |
| L90 | 16 |
| # N's per 100 kbp | 1.13 |

**Table 1.** Genome assembly metrics for *Argopecten irradians* NY.

length of 14,207 bp. The flash-frozen adductor muscle was processed for Hi-C library construction using an Arima Genomics Hi-C Kit (San Diego, California, USA) according to the manufacturer's guidelines. This Hi-C library was then sequenced on a single lane of an Illumina HiSeqX PE150, resulting in a total of 779,291,520 paired-end reads.

**Transcriptome sequencing.** Transcriptomic data were generated from kidney samples derived from a total of 137 wild and aquacultured scallops collected from Orient Harbor, New York, and used in laboratory experiments or deployed in Flanders Bay (40.917634, −72.593486). RNA was extracted using the NucleoSpin® RNA Plus RNA isolation kit (Macherey-Nagel, Düren, Germany). RNA quantity and quality were checked spectrophotometrically (NanoDrop® ND-1000, Thermo Fisher Scientific, Wilmington, Delaware, USA). Library preparation and sequencing were performed by Novogene Corporation (UC Davis, Sacramento, California, USA). Sample quality control measures implemented by Novogene rely mainly on RNA Nano 6000 Assay Kit using the Bioanalyzer 2100 system (Agilent Technologies, Santa Clara, California, USA). RNA-seq libraries were prepared using 1 µg RNA using NEBNext Ultra RNA Library Prep Kit for Illumina (New England Biolabs, Ipswich, Massachusetts USA). Library fragments were purified with AMPure XP system (Beckman Coulter, Beverly, California, USA). Then 3 µl USER Enzyme (New England Biolabs, USA) was used with size-selected, adaptor-ligated cDNA at 37 °C for 15 min followed by 5 min at 95 °C before PCR. Then PCR was performed with Phusion High-Fidelity DNA polymerase, Universal PCR primers and Index (X) Primer. Finally, PCR products were purified (AMPure XP system) and library quality was assessed on the Agilent Bioanalyzer 2100 system. The clustering of the index-coded samples was performed on a cBot Cluster Generation System using PE Cluster Kit cBot-HS (Illumina) before sequencing on an Illumina platform where 150 bp paired-end reads were generated.

**Genome assembly.** The initial assembly was generated from sequences derived from all PacBio reads after adaptor removal using BBmap's removesmartbell.sh script. Strategies recommended by Guiglielmoni *et al.*[30] were adopted, using the Raven assembler[31] with default parameters to produce a 1Gb-size assembly. Potential uncollapsed haplotypes were removed using Purge Haplotigs[32]. A polishing process was then conducted using HyPo[33]. Scaffolding of this assembly was further achieved using Hi-C data. Hi-C reads were processed with hicstuff[34] with the parameters --enzyme DpnII,HinfI --iterative. This processing pipeline incorporated a mapping step against the contigs using Bowtie 2[35]. instaGRAAL[36] was run with --level 5 --cycles 100 --coverage-std 1 --neighborhood 5 parameters, with further automatic curation from instagraal-polish script. Blobtools[37] was run with default parameters on the final scallop assembly to detect potential contamination. For this, Illumina reads were mapped on the assembly using the BWA mem algorithm[38] and BLASTn v. 2.11.0[39] was run against the NT database from NCBI[40], providing input to Blobtools. This workflow generated a chromosome-level genome assembly of the bay scallop that contains a total of 845.9 Mb distributed over 1,503 scaffolds with a GC content of 35.6%. The scaffolds have an $N_{50}$ of 44.3 Mb ($L_{50}$ = 8 scaffolds) and an $N_{90}$ of 34.5 Mb ($L_{90}$ = 16 scaffolds) (Table 1). Confirmatory Hi-C analysis revealed the presence of 16 chromosome pairs in *A. irradians* (Fig. 2).

The majority (92.9%) of our assembled genome is contained within the 16 largest scaffolds which ranges from 75.5 Mb to 34.5 Mb. In addition to the nuclear genome, the complete mitochondrial genome of 16,414 bp was successfully assembled.

**Genome annotation.** RepeatModeler v. 2.0.4[41] was used to identify repetitive elements in the genome of *A. irradians*. Tandem repeats were identified using Tandem Repeats Finder v. 4.0.10[42] with recommended parameters. Repeats and low-complexity DNA sequences were masked using RepeatMasker v. 4.1.5[43]. The repeat content of the *A. irradians* genome (Table 2) is similar to those reported in *Pecten maximus*[23] and *Mizuhopecten yessoensis*[20]. Total interspersed repeats represent 36.2% of the *A. irradians* genome, which is closer to the 38.9% observed in *M. yessoensis* and higher than the 25.8% seen in *P. maximus*.

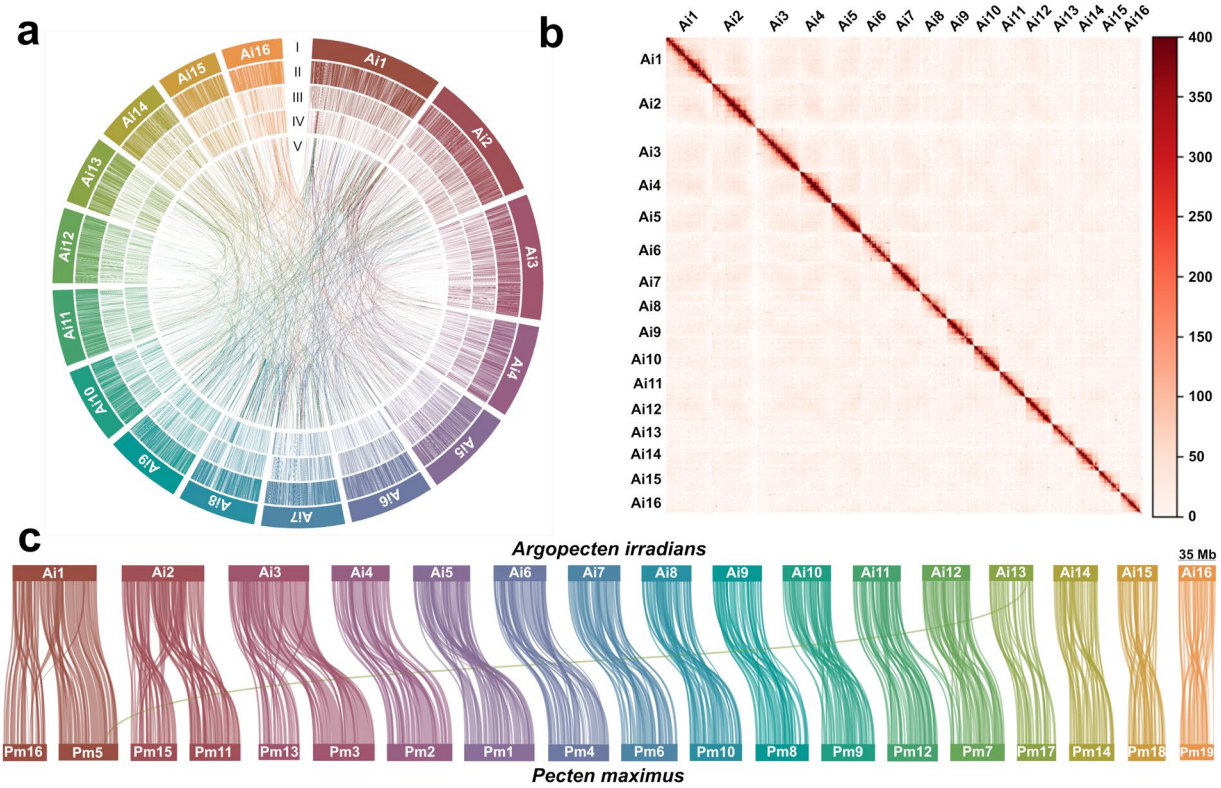


**Fig. 2** Chromosomal organization and synteny in *Argopecten irradians* NY. (**a**) A circular genomic map illustrating 16 chromosomes with color-coded segments (I), detailing GC content (II), gene (III) and repeated sequence density (IV). Interconnecting lines (V) represent syntenic blocks and conserved genomic regions across the chromosomes. (**b**) Contact map of the *A. irradians* genome assembly. Map generated from Hi-C data showing sequence interaction points in chromosomes (red dots). The color bar indicates contact density. (**c**) Genomic collinearity between *A. irradians* and the king scallop, *P. maximus*.

Prediction of protein-coding genes was based on *ab initio* gene predictions, homology-based predictions, and transcriptome-based predictions. *Ab initio* predictions were performed by Augustus v. 3.5[44], GlimmerHMM v. 3.0.2[45], and SNAP[46]. For homology-based prediction, GeMoMa v. 1.9[47] was used to annotate the gene models in *A. irradians* NY using amino acid sequences from *P. maximus*[23], *M. yessoensis*[20], *A. irradians* (subspecies *irradians* and *concentricus*)[25] genomes and TOGA[48] was used with the human genome (hg38) as the reference. For RNA-seq-based prediction, the clean RNA-seq reads were aligned to the assembled genome using HISAT2 v2.2.1[49] and were assembled by StringTie v. 2.2.0[50] with the default parameters, and then TransDecoder v5.5.0 (https://github.com/TransDecoder/TransDecoder) and PASA v. 2.4.1[51] were jointly used for final coding-gene prediction. All gene structures predicted by the above methods were integrated into a nonredundant gene set using EVidenceModeler v. 1.1.1[51]. The weight value was set to 10 for high-quality RNA-seq transcripts, 5 for high-quality homologous proteins, and 1 for ab initio predicted transcripts. Finally, the resulting protein models were finally functionally annotated by integrating the annotation information from InterProScan v. 5.63–95.0[52], KOALA (KEGG Orthology And Links Annotation)[53], and the eggNOG-mapper v. 2.0.1[54,55]. Noncoding RNA was annotated using RNAmmer v. 1.2[56] for rRNA, tRNAscan-SE v. 2.0[57] for tRNA and the cmscan module in Infernal v. 1.1.2[58] for miRNA, snRNA and snoRNA. A comprehensive annotation of protein-coding sequences was achieved through a multifaceted approach that integrated *de novo* gene prediction, protein homology searches, and transcriptome-based predictions. This analysis allowed the identification of 33,772 genes with an average length of 8,563 bp. The mean coding sequence length was 1,382 bp, with an average of 6.14 exons per gene and an average exon length of 225 bp.

Our comparative genomic analysis considered another scallop species for which a high-quality genome assembly exists (king scallop) and revealed a significant structural divergence between the *A. irradians* and *P. maximus* genomes, highlighting a pattern consistent with chromosomal fusion events. Our results revealed the presence of 16 chromosome pairs in *A. irradians*, consistent with previous karyotype evidence[59]. These findings support chromosomal rearrangements and fusions. For instance, scaffolds Ai1, Ai2, and Ai3 *of A. irradians* exhibit syntenic blocks that align with several chromosome-scale scaffolds of *P. maximus* (Fig. 2c), supporting the hypothesis that these chromosomes are products of an ancestral chromosomal fusion[60]. This finding is consistent with the observed reduction in chromosome number from the ancestral 19[20], aligning with previous studies on chromosomal evolution within the Pectinidae[61]. The estimated time of divergence of *A. irradians* and *A. purpuratus* ~14 million years ago is consistent with fossil data which suggests their separation occurred during the Miocene epoch[62]. Notably, *A. irradians*[59,63] and *A. purpuratus* both exhibit a haploid

| Element Type | Number of Elements | Length Occupied (bp) | Percentage of Sequence |
|---|---|---|---|
| Retroelements | 111,065 | 45,503,269 | 5.43 |
| SINEs: | 11,701 | 1,856,540 | 0.22 |
| Penelope: | 0 | 0 | 0 |
| LINEs: | 58,361 | 25,575,385 | 3.05 |
| L2/CR1/Rex | 12,228 | 4,536,621 | 0.54 |
| R1/LOA/Jockey | 5,537 | 4,754,207 | 0.57 |
| R2/R4/NeSL | 2,796 | 1,345,314 | 0.16 |
| RTE/Bov-B | 8,562 | 3,761,032 | 0.45 |
| L1/CIN4 | 290 | 306,683 | 0.04 |
| LTR elements: | 41,003 | 18,071,344 | 2.16 |
| BEL/Pao | 439 | 604,652 | 0.07 |
| Ty1/Copia | 857 | 245,407 | 0.03 |
| Gypsy/DIRS1 | 10,155 | 11,459,235 | 1.37 |
| Retroviral | 566 | 281,038 | 0.03 |
| DNA transposons | 34,399 | 7,180,469 | 0.86 |
| hobo-Activator | 2,584 | 714,065 | 0.09 |
| Tc1-IS630-Pogo | 21,349 | 3,210,208 | 0.38 |
| MULE-MuDR | 462 | 240,239 | 0.03 |
| PiggyBac | 223 | 156,048 | 0.02 |
| Tourist/Harbinger | 536 | 202,879 | 0.02 |
| Other (Mirage, P-element, Transib) | 839 | 117,152 | 0.01 |
| Rolling-circles | 6,338 | 647,513 | 0.08 |
| Unclassified | 1,121,661 | 250,565,449 | 29.89 |
| Total interspersed | | 303,249,187 | 36.18 |
| Small RNA | 13,500 | 2,191,028 | 0.26 |
| Satellites | 1,888 | 562,954 | 0.07 |
| Simple repeats | 91,158 | 4,939,239 | 0.59 |
| Low complexity | 17,427 | 851,457 | 0.1 |

**Table 2.** The interspersed repeat content of the *Argopecten irradians* NY genome.

number of 16 chromosomes, deviating from the ancestral state and indicating a lineage-specific reduction. The selective advantage of chromosomal fusions, such as the creation of new gene linkages or the loss of redundant genetic material, is consistent with the concept of local adaptation and the evolution of chromosome fusions[64].

**Phylogenetic analysis and divergence time estimation.** The genome of *A. irradians* NY and ten other molluscan genomes were used for gene family construction using OrthoFinder v. 2.5.5[65] with default parameters. The protein sequences of 281 single-copy orthologs from 11 species were independently aligned using MUSCLE[66], curated using Gblocks v. 0.91b[67] with an option to allow gap positions within the final blocks, and then concatenated using PhyloSuite v. 1.2.2[68] for species tree construction. The maximum likelihood tree was calculated using IQ-TREE[69], based on the recommendations of ModelFinder[70], and branching support was estimated using UFBoot[71]. BEAST 2 v. 2.7.5[72] was used to estimate species divergence times with the JTT substitution model and gamma categories equal to 4. The calibrated Yule model and strict clock type were set. The chain length for MCMC was set to 10,000,000 and the parameters were recorded every 1,000 generations. The calibration points used in BEAST 2 were obtained from the TimeTree database[73]: *Octopus bimaculoides* versus Bivalvia (median time: 520 MYA), *Crassostrea virginica* versus *Magallana gigas* (median time: 73 MYA). The gene-family expansion and contraction were determined using CAFE5[74]. The gene family size for each species used in CAFE was calculated by OrthoFinder v. 2.5.5[65]. Comparative genomic analysis of 11 molluscan species, including *A. irradians*, has revealed major evolutionary events and gene expansions and contractions (Fig. 3). The phylogenetic timeline derived from shared gene sets estimates the divergence of *A. irradians* and *A. purpuratus* between 12.6 and 15.9 million years ago (Fig. 3a). Gene clustering analysis using OrthoFinder revealed 7,036 gene families shared among all analyzed molluscan genomes. Among shared gene families, a higher occurrence of shared genes was represented in only one copy within the genus *Argopecten*, where 84.3–84.4% of the shared gene families were single-copy. It was also found that 1,168 genes in *A. irradians* and 858 genes in *A. purpuratus* were clustered into 640 gene families exclusive to *Argopecten* species and 1792 gene families found in Pectinidae species (Fig. 3b). The *A. irradians* genome exhibits an expansion of 420 gene families. Enrichment analysis of these families reveals the most substantial increase noted in the phagosome pathway (Fig. 3c).

Enriched pathways include amino sugar and nucleotide sugar metabolism, glycosaminoglycan biosynthesis - chondroitin sulfate, and the phosphatidylinositol signaling system. Additionally, we observed significant enrichment in metabolic pathways, such as fatty acid metabolism, arachidonic acid metabolism, and drug metabolism involving cytochrome P450 enzymes.

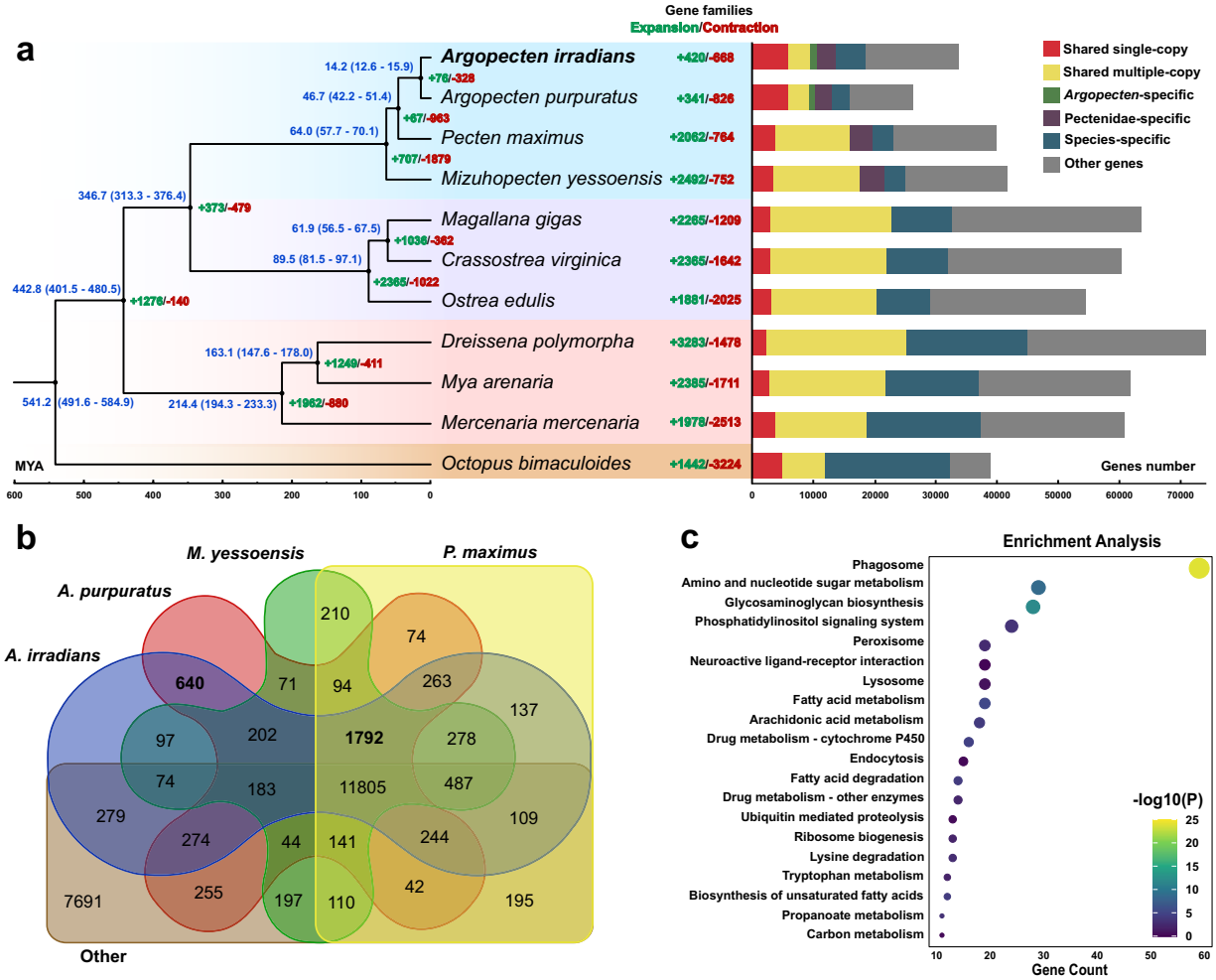


**Fig. 3** Gene family dynamics and functional enrichment in molluscan evolution. (**a**) Phylogenetic analysis and gene family evolution of 11 molluscan species. Tree topology with estimated divergence times (million years ago, MYA, including range) is shown next to each lineage (blue). Number of expanded (green) and contracted (red) gene families shown next each branch. The right panel illustrates the distribution of gene families among 11 molluscan species. It shows the number of shared single-copy genes (red), which are present in all analyzed genomes but exist in only one copy per individual genome; shared multiple-copy genes (yellow), which are present in all analyzed genomes and exist in multiple copies in individual genomes; genes found only in *Argopecten* species (green); genes found in Pectinidae species (purple); genes found only in the corresponding species (blue); and other species-combination of orthologous genes (grey). (**b**) Venn Diagram graph of orthologous gene families shared/not shared among 11 molluscan species. (**c**) Enriched KEGG pathway analysis for expanded gene families in *Argopecten irradians*. The bubble plot illustrates the top 20 enriched pathways, where the bubble size reflects the number of genes and the color gradient ($-\log_{10}$(P-value)) indicates the significance of the enrichment.

## Data Records

The raw sequencing data and genome assembly *of A. irradians* have been deposited at the National Center for Biotechnology Information (NCBI) under BioProject PRJNA1050236. The assembled genome has been deposited in the NCBI assembly with the accession number JAYEEO000000000[75]. The raw PacBio, Illumina Hi-C, and transcriptome data have been deposited in the Sequence Read Archive (SRA) repository with the accession number of SRP478220[76]. Additionally, the results of annotation have been deposited in the Figshare[77] and Dryad[78] databases.

## Technical Validation

Quality of the final assembly was evaluated using the Benchmarking Universal Single-Copy Orthologs (BUSCO v. 5.3.0)[79] analysis with the Metazoa_odb10 lineage. We found 96.2% complete (among which 1.9% are duplicated), and 98.3% complete + fragmented, BUSCO core genes represented in the Metazoa (odb10) BUSCO database (Table 3). Additionally, BUSCO analysis was performed on the annotated proteins, yielding 94.8% complete BUSCOs (of which 2.8% are duplicated), further supporting the quality of the genome annotation.

| BUSCO analysis | Genome assembly | Annotated proteins |
|---|---|---|
| Complete BUSCOs (C) | 918 (96.2%) | 905 (94.8%) |
| Complete and single-copy BUSCOs (S) | 900 (94.3%) | 878 (92.0%) |
| Complete and duplicated BUSCOs (D) | 18 (1.9%) | 27 (2.8%) |
| Fragmented BUSCOs (F) | 20 (2.1%) | 29 (3.0%) |
| Missing BUSCOs (M) | 16 (1.7%) | 20 (2.2%) |
| Total BUSCO groups searched | 954 (100%) | 954 (100%) |

**Table 3.** BUSCO analysis of genome assembly and annotated proteins completeness for *Argopecten irradians* NY.

The high quality of the genome assembly is demonstrated by the successful mapping of 96.94% ± 1.61% of transcriptomic reads, as well as second and third generation sequencing data, with mapping rates of 97.72% and 95.87%, respectively (Supplementary Table 1).

## Code availability

All analyses followed the guidelines provided in the manuals and tutorials for the software and pipeline used. The specific software versions used are detailed in the Methods section. Default settings or those recommended by the authors were used for the software and analysis pipeline, unless otherwise noted.

Received: 10 June 2024; Accepted: 19 September 2024;
Published online: 28 September 2024

## References

1. Adamkewicz, S. L., Harasewych, M. G., Blake, J., Saudek, D. & Bult, C. J. A molecular phylogeny of the bivalve mollusks. *Molecular Biology and Evolution* **14**, 619–629 (1997).
2. Cummings, K. S. & Graf, D. L. Mollusca. in *Ecology and Classification of North American Freshwater Invertebrates* 309–384 (Elsevier, 2010).
3. Strehse, J. S. & Maser, E. Marine bivalves as bioindicators for environmental pollutants with focus on dumped munitions in the sea: A review. *Marine Environmental Research* **158**, 105006 (2020).
4. Chahouri, A., Yacoubi, B., Moukrim, A. & Banaoui, A. Bivalve molluscs as bioindicators of multiple stressors in the marine environment: Recent advances. *Continental Shelf Research* **264**, 105056 (2023).
5. Jørgensen, C. Bivalve filter feeding revisited. *Mar Ecol Prog Se.* **142**, 287–302 (1996).
6. *The State of World Fisheries and Aquaculture 2022. Towards Blue Transformation*. (FAO, Rome, 2022).
7. Minchin, D. Introductions: some biological and ecological characteristics of scallops. *Aquatic Living Resources* **16**, 521–532 (2003).
8. Zhan, A. *et al*. Fine-scale population genetic structure of Zhikong scallop (*Chlamys farreri*): do local marine currents drive geographical differentiation? *Mar Biotechnol* **11**, 223–235 (2009).
9. Adamkewicz, L. & Castagna, M. Genetics of shell color and pattern in the bay scallop *Argopecten irradians*. *Journal of Heredity* **79**, 14–17 (1988).
10. Estabrooks, S. L. The possible role of telomeres in the short life span of the bay scallop, *Argopecten irradians irradians* (Lamarck 1819). *Journal of Shellfish Research* **26**, 307–313 (2007).
11. Guderley, H. E. & Tremblay, I. Swimming in scallops. in *Developments in Aquaculture and Fisheries Science* vol. 40 535–566 (Elsevier, 2016).
12. Bert, T. M., Arnold, W. S., McMillen-Jackson, A. L., Wilbur, A. E. & Crawford, C. Natural and anthropogenic forces shape the population genetics and recent evolutionary history of eastern United States bay scallops (*Argopecten irradians*). *Journal of Shellfish Research* **30**, 583–608 (2011).
13. Waller, T. R. The evolution of the *Argopecten gibbus* stock (Mollusca: Bivalvia), with emphasis on the tertiary and quaternary species of Eastern North America. *J. Paleontol.* **43**, 1–125 (1969).
14. Bologna, P., Wilbur, A. E. & Able, K. Reproduction, population structure, and recruitment limitation in a bay scallop (*Argopecten irradians* Lamarck) population from New Jersey, USA. *Journal of Shellfish Research* **20**, 89–96 (2001).
15. Fusui, Z. *et al*. Introduction, spat-rearing and experimental culture of bay scallop, *Argopecten irradians* Lamarck. *Chin. J. Ocean. Limnol.* **9**, 123–131 (1991).
16. Yu, L. *et al*. Value chain of the data-poor Chinese bay scallop aquaculture. *Marine Policy* **150**, 105556 (2023).
17. Guo, X. & Luo, Y. Scallops and scallop aquaculture in China. in *Developments in Aquaculture and Fisheries Science* vol. 40 937–952 (Elsevier, 2016).
18. Zhang, G. *et al*. The oyster genome reveals stress adaptation and complexity of shell formation. *Nature* **490**, 49–54 (2012).
19. Sun, J. *et al*. Adaptation to deep-sea chemosynthetic environments as revealed by mussel genomes. *Nat Ecol Evol* **1**, 0121 (2017).
20. Wang, S. *et al*. Scallop genome provides insights into evolution of bilaterian karyotype and development. *Nat Ecol Evol* **1**, 120 (2017).
21. Li, Y. *et al*. Scallop genome reveals molecular adaptations to semi-sessile life and neurotoxins. *Nat Commun* **8**, 1721 (2017).
22. Li, C. *et al*. Draft genome of the Peruvian scallop *Argopecten purpuratus*. *GigaScience* **7**, (2018).
23. Kenny, N. J. *et al*. The gene-rich genome of the scallop *Pecten maximus*. *GigaScience* **9**, giaa037 (2020).
24. Fletcher, C. *et al*. The genome sequence of the variegated scallop, *Mimachlamys varia* (Linnaeus, 1758). *Wellcome Open Res* **8**, 307 (2023).
25. Liu, X. *et al*. Draft genomes of two Atlantic bay scallop subspecies *Argopecten irradians irradians* and *A. i. concentricus*. *Sci Data* **7**, 99 (2020).
26. Wang, L., Zhang, H., Song, L. & Guo, X. Loss of allele diversity in introduced populations of the hermaphroditic bay scallop *Argopecten irradians*. *Aquaculture* **271**, 252–259 (2007).
27. Pales Espinosa, E. *et al*. An apicomplexan parasite drives the collapse of the bay scallop population in New York. *Sci Rep* **13**, 6655 (2023).
28. Mathur, V. *et al*. Phylogenomics identifies a new major subgroup of Apicomplexans, Marosporida *class nov*., with extreme apicoplast genome reduction. *Genome Biology and Evolution* **13**, evaa244 (2021).
29. Sambrook, J., Fritsch, E. F., Maniatis, T., Russell, D. W. & Green, M. R. *Molecular Cloning: A Laboratory Manual*. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1989).
30. Guiglielmoni, N., Houtain, A., Derzelle, A., Van Doninck, K. & Flot, J.-F. Overcoming uncollapsed haplotypes in long-read assemblies of non-model organisms. *BMC Bioinformatics* **22**, 303 (2021).

31. Vaser, R. & Šikić, M. Time- and memory-efficient genome assembly with Raven. *Nat Comput Sci* **1**, 332–336 (2021).
32. Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* **19**, 460 (2018).
33. Kundu, R., Casey, J. & Sung, W.-K. *HyPo: Super Fast & Accurate Polisher for Long Read Genome Assemblies*. https://doi.org/10.1101/2019.12.19.882506 (2019).
34. Matthey-Doret, C. *et al*. Computer vision for pattern detection in chromosome contact maps. *Nat Commun* **11**, 5795 (2020).
35. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012).
36. Baudry, L. *et al*. instaGRAAL: chromosome-level quality scaffolding of genomes using a proximity ligation-based scaffolder. *Genome Biol* **21**, 148 (2020).
37. Laetsch, D. R. & Blaxter, M. L. BlobTools: Interrogation of genome assemblies. *F1000Res* **6**, 1287 (2017).
38. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
39. Camacho, C. *et al*. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
40. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **33**, D501–504 (2005).
41. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA* **117**, 9451–9457 (2020).
42. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**, 573–580 (1999).
43. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* Chapter 4, 4.10.1-4.10.14 (2009).
44. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research* **34**, W435–W439 (2006).
45. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
46. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
47. Keilwagen, J. *et al*. Using intron position conservation for homology-based gene prediction. *Nucleic Acids Res* **44**, e89–e89 (2016).
48. Kirilenko, B. M. *et al*. Integrating gene annotation with orthology inference at scale. *Science* **380**, eabn3107 (2023).
49. 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 (2019).
50. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290–295 (2015).
51. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9**, R7 (2008).
52. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
53. Kanehisa, M., Sato, Y. & Morishima, K. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J Mol Biol* **428**, 726–731 (2016).
54. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Molecular Biology and Evolution* **38**, 5825–5829 (2021).
55. Huerta-Cepas, J. *et al*. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research* **47**, D309–D314 (2019).
56. Lagesen, K. *et al*. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research* **35**, 3100–3108 (2007).
57. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research* **25**, 955–964 (1997).
58. Nawrocki, E. P., Kolbe, D. L. & Eddy, S. R. Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25**, 1335–1337 (2009).
59. Huang, X. *et al*. Cytogenetic characterization of the bay scallop, *Argopecten irradians irradians*, by multiple staining techniques and fluorescence *in situ* hybridization. *Genes Genet Syst* **82**, 257–263 (2007).
60. Wang, Y. & Guo, X. Chromosomal rearrangement in Pectinidae revealed by rRNA loci and implications for bivalve evolution. *The Biological Bulletin* **207**, 247–256 (2004).
61. Zhang, L., Bao, Z., Wang, S., Huang, X. & Hu, J. Chromosome rearrangements in Pectinidae (Bivalvia: Pteriomorphia) implied based on chromosomal localization of histone H3 gene in four scallops. *Genetica* **130**, 193–198 (2007).
62. Waller, T. R. The evolution of the *Argopecten gibbus* stock (Mollusca: Bivalvia), with emphasis on the tertiary and quaternary species of Eastern North America. *Memoir (The Paleontological Society)* **3**, i-v+1–125 (1969).
63. Gajardo, G., Parraguez, M. & Colihueque, N. Karyotype analysis and chromosome banding of the Chilean-Peruvian scallop *Argopecten purpuratus* (Lamarck, 1819). *J. Shellfish Res* **21**, 585–590 (2002).
64. Guerrero, R. F. & Kirkpatrick, M. Local adaptation and the evolution of chromosome fusions. *Evolution* **68**, 2747–2756 (2014).
65. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* **16**, 157 (2015).
66. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792–1797 (2004).
67. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution* **17**, 540–552 (2000).
68. Zhang, D. *et al*. PhyloSuite: An integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies. *Molecular Ecology Resources* **20**, 348–355 (2020).
69. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**, 268–274 (2015).
70. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* **14**, 587–589 (2017).
71. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol Biol Evol* **35**, 518–522 (2018).
72. Bouckaert, R. *et al*. BEAST 2: A software platform for Bayesian evolutionary analysis. *PLoS Comput Biol* **10**, e1003537 (2014).
73. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: A resource for timelines, timetrees, and divergence times. *Molecular Biology and Evolution* **34**, 1812–1819 (2017).
74. Mendes, F. K., Vanderpool, D., Fulton, B. & Hahn, M. W. CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics* **36**, 5516–5518 (2021).
75. *NCBI Nucleotide*. http://identifiers.org/nucleotide:JAYEEO000000000.1 (2024).
76. *NCBI Sequence Read Archive*. http://identifiers.org/insdc.sra:SRP478220 (2024).
77. Grouzdev, D. *et al*. Chromosome-level genome assembly of the bay scallop *Argopecten irradians*. *Figshare* https://doi.org/10.6084/m9.figshare.27015544 (2024).
78. Grouzdev, D. *et al*. Chromosome-level genome assembly of the bay scallop *Argopecten irradians*. *Dryad* https://doi.org/10.5061/dryad.d51c5b09b (2024).
79. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: Assessing genome assembly and annotation completeness. *Methods Mol Biol* **1962**, 227–245 (2019).

## Acknowledgements
This work was supported by NSF Grant number IOS-2026358 to B.A., E.P.E. and S.T. Financial support was also provided by the McConnell Family Foundation (B.A., E.P.E., D.G.) and by the New York State Department of Environmental Conservation (B.A., E.P.E.). N.G. and J-F.F. were funded by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 764840.

## Author contributions
E.P.E., B.A. and S.T. designed the study and secured the funding. S.T., H.T. and E.P.E. collected and processed biological samples. A.T. and I.B. generated Hi-C libraries. N.G. and J.-F.F. assisted with Hi-C data analysis. D.G., S.F. performed genomic analysis. D.G., E.P.E. and B.A. analyzed data and drafted the paper. All authors contributed to the editing of the manuscript and approved the final version of the paper.

## Competing interests
The authors declare no competing interests.

## Additional information
**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41597-024-03904-x.

**Correspondence** and requests for materials should be addressed to B.A.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.