






OPEN

DATA DESCRIPTOR

# Chromosome-level genome assembly of the northern Pacific seastar *Asterias amurensis*

Yanlin Wang<sup>1</sup> , Yixin Wang<sup>1</sup>, Yujia Yang<sup>1</sup>, Gang Ni<sup>1</sup>, Yulong Li<sup>2</sup>  & Muyan Chen<sup>1</sup> 

*Asterias amurensis* has attracted widespread concern because of its population outbreaks, which has impacted fisheries and aquaculture, as well as disrupting local ecosystems. A high-quality reference genome is necessary to better investigate mechanisms of outbreak and adaptive changes. Combining PacBio HiFi and Hi-C sequencing data, we generated a chromosome-level *A. amurensis* genome with a size of 491.53 Mb. The contig N50 and scaffold N50 were 8.05 and 23.75 Mb, respectively. The result of BUSCO analysis revealed a completeness score of 98.85%. A total of 16,531 protein-coding genes were predicted in the genome, of which 94.63% were functionally annotated. The high-quality genome assembly resulting from this study will provide a valuable genetic resource for future research on the mechanism of population outbreaks and invasion ecology.

## Background & Summary

*Asterias amurensis* (class: Asterozoa), also known as the northern Pacific seastar, is widely distributed in the northwest and northeast Pacific, native to the coast of Alaska<sup>1</sup>, China<sup>2</sup>, Japan<sup>3</sup>, Korea<sup>4</sup>, and Russia<sup>5</sup>. As a benthic echinoderm with distinct evolutionary classification<sup>6</sup>, its reproduction mode includes not only dioecious but also asexual reproduction by arm regeneration<sup>7,8</sup>. Females have high fecundity and can annually spawn ~20 million eggs<sup>3</sup>. The planktonic stage of larva can last for seven weeks or several months, which enables them to rapidly spread in a suitable environment<sup>9,10</sup>. *A. amurensis* is located at the highest trophic level among the benthic invertebrates as a voracious and efficient generalist predator<sup>11</sup>, which has been reported to impact a variety of infaunal taxa, especially commercial bivalves<sup>12–14</sup>. And it has even been associated with the decline of some fish species<sup>15</sup>.

In the early 1980s, free-spawning starfish *A. amurensis* were first spotted in southeast Tasmania of Australia, possibly introduced from central Japan through ship ballast water<sup>3</sup>. Since their first detection, this starfish has successfully established populations in a short period and gradually expanded to Victoria<sup>16–18</sup>. As one of the most successful invasive species, *A. amurensis* became a significant threat to native assemblages, marine commercial species, and has damaged native ecosystems in Australia<sup>13,19</sup>. Thus, this starfish was listed as one of the high-priority marine pests in Australia<sup>20</sup>. Although its invasive range is limited in Australia<sup>21</sup> so far, *A. amurensis* will likely continue to expand due to its high fecundity, wide environmental tolerance, and long larval duration<sup>22</sup>, even invading the Southern Ocean<sup>23</sup>. However, due to the lack of genomic information in *A. amurensis*, genetic changes associated with invasive lineages remain unknown<sup>16,24</sup>.

Periodic and massive outbreaks of *A. amurensis* populations have been reported in several countries, including Australia, China, and Japan, which have significantly impacted fishery and mariculture grounds, as well as destroyed the original ecological balance, leading to serious economic losses<sup>25–27</sup>. Unfortunately, no effective bio-control method has been reported for this pest up to now. To provide warning information for possible outbreaks of *A. amurensis*, early detection technologies have been developed based on targeting rRNA<sup>28</sup> and the mitochondrial cytochrome c oxidase subunit I (COI) gene<sup>21,29,30</sup>. However, the mechanism of aggregation and outbreak is complex and unclear. Relevant studies require the support of a high-quality genome assembly, which may help to identify species-specific factors associated with aggregating starfish<sup>31</sup>.

<sup>1</sup>The Key Laboratory of Mariculture, Ministry of Education, Ocean University of China, Qingdao, 266003, China.

<sup>2</sup>CAS Key Laboratory of Marine Ecology and Environmental Sciences, Institute of Oceanology, Chinese Academy of Sciences, Qingdao, 266071, China. ✉e-mail: [lyl@qdio.ac.cn](mailto:lyl@qdio.ac.cn); [chenmuyan@ouc.edu.cn](mailto:chenmuyan@ouc.edu.cn)

Libraries	Insert size (bp)	Clean data (Gb)	Reads number	Read length (bp)	Sequence coverage (X)
BGI reads	350	112.58	377,205,535	150	229.04
PacBio reads	15,000	11.15	890,929	12,511 (mean)	22.68
RNA-seq	350	13.47	45,079,838	150	—
Total	—	137.20	423,176,302	—	251.72

**Table 1.** Statistical analysis of sequencing reads from BGI, Illumina and PacBio.

Type	Data
Raw paired reads	350,304,882
Raw Base(bp)	105,091,464,600
Clean Base(bp)	102,748,760,698
Effective Rate(%)	99.77
Q20(%)	97.32
Q30(%)	92.33
GC Content(%)	39.18

**Table 2.** Statistical analysis of sequencing data from Hi-C.

Type	Contig (bp)	Scaffold (bp)
Total Number	90	22
Total Length	491,503,102	491,537,102
Average Length	5,461,145	22,342,596
Max Length	28,598,918	38,009,675
N50 Length	8,054,564	23,750,475
N50 Number	19	9
N90 Length	3,441,115	15,767,093
N90 Number	54	19

**Table 3.** Assembly statistics of *A. amurensis* genome.

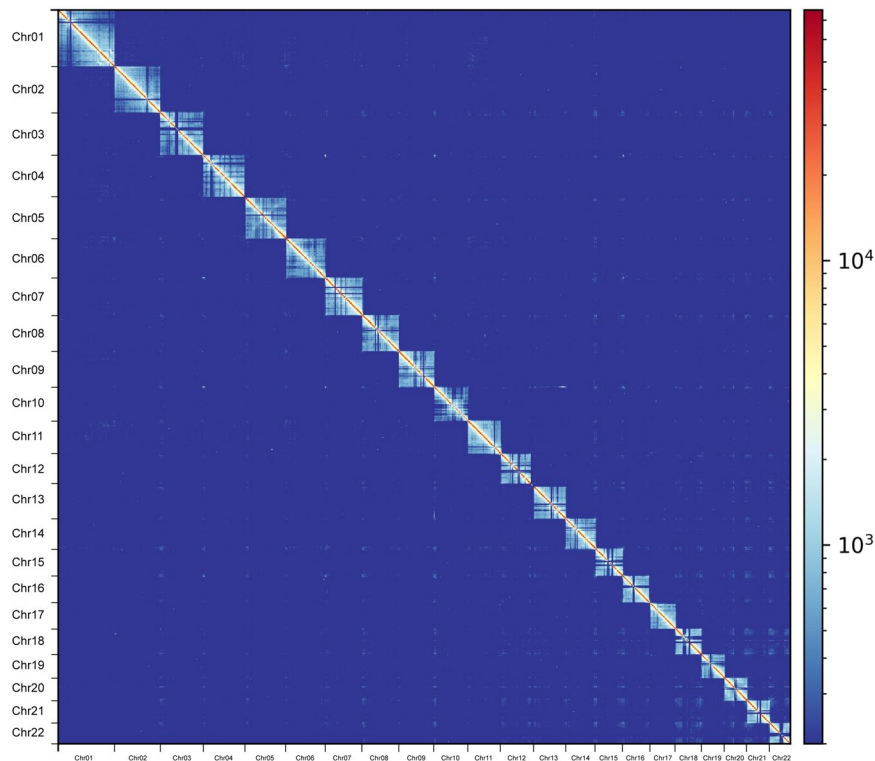
In the present study, a *de novo* assembled chromosome-level *A. amurensis* genome was prepared using PacBio HiFi and Hi-C sequencing data. The final genome size was 491.53 Mb with scaffold N50 of 23.75 Mb. Using three approaches for gene structure annotation, we identified a total of 16,531 protein-coding genes, of which 15,643 genes were functionally annotated with at least one public database. A high-quality reference genome for *A. amurensis* will be a useful genomic resource to explore both the mechanism of population outbreak and the genetic basis underlying adaptive change during the invasion process. Meanwhile, the *A. amurensis* genome will be a noteworthy addition to the existing suite of Asteroidea genomes for future cell, developmental and evolutionary biology research.

## Methods

**Sample collection.** All samples used in this study were from a male adult *A. amurensis* collected by diving in Qingdao, Shandong Province, China (36°03'04"N, 120°21'26"E) in November 2022. Fresh gonad tissue from the base of the arm was excised and washed with phosphate buffered saline (PBS, 1X). It was then immediately frozen in liquid nitrogen and transferred to  $-80^{\circ}\text{C}$  for storage. High quality DNA was extracted from gonad using DNeasy Blood & Tissue Kit (Qiagen, Germany) for long-read and short-read whole genome sequencing. To aid in structural annotation, nine tissues including gonad, body wall, madreporite, spine, mouth, stomach, muscle, podia, and eye spot were used for transcriptome sequencing. All tissues were isolated separately with scissors and forceps, and then treated in the same way as the gonad collection. Total RNA was extracted using the TRIzol reagent (Vazyme, China).

**Sequencing.** For long-read sequencing, high molecular weight genomic DNA (gDNA) was fragmented to approximately 15 kb to construct a PacBio HiFi library. The sequencing library was generated using the SMRTbell Express Template Prep kit 2.0 (Pacific Biosciences, USA), following the manufacturer's recommendations, as described in the previous study<sup>32</sup>. The library was finally sequenced with circular consensus sequencing (CCS) mode on the PacBio Sequel II system using a single 8 M cell. After filtering out the low-quality reads and sequence adapters, a total of 11.15 Gb CCS data were obtained with a mean length of 12.51 kb (Table 1).

For short-read whole genome sequencing, gDNA was fragmented into approximately 350 bp for library construction. The library was sequenced on DNBSEQ-T7 platform to generate 150 bp paired-end (PE150) reads. After filtering out low-quality reads including reads shorter than 100 bp, reads that contained >10% "N", and



**Fig. 1** Genome-wide heatmap of Hi-C interactions among 22 chromosomes in *A. amurensis*. The scale bar represents the interaction frequency of Hi-C links.

reads that contained >50% low-quality bases (Phred score  $\leq 10$ ), the clean data generated was 112.58 Gb, which covered  $\sim 229\times$  of the genome (Table 1).

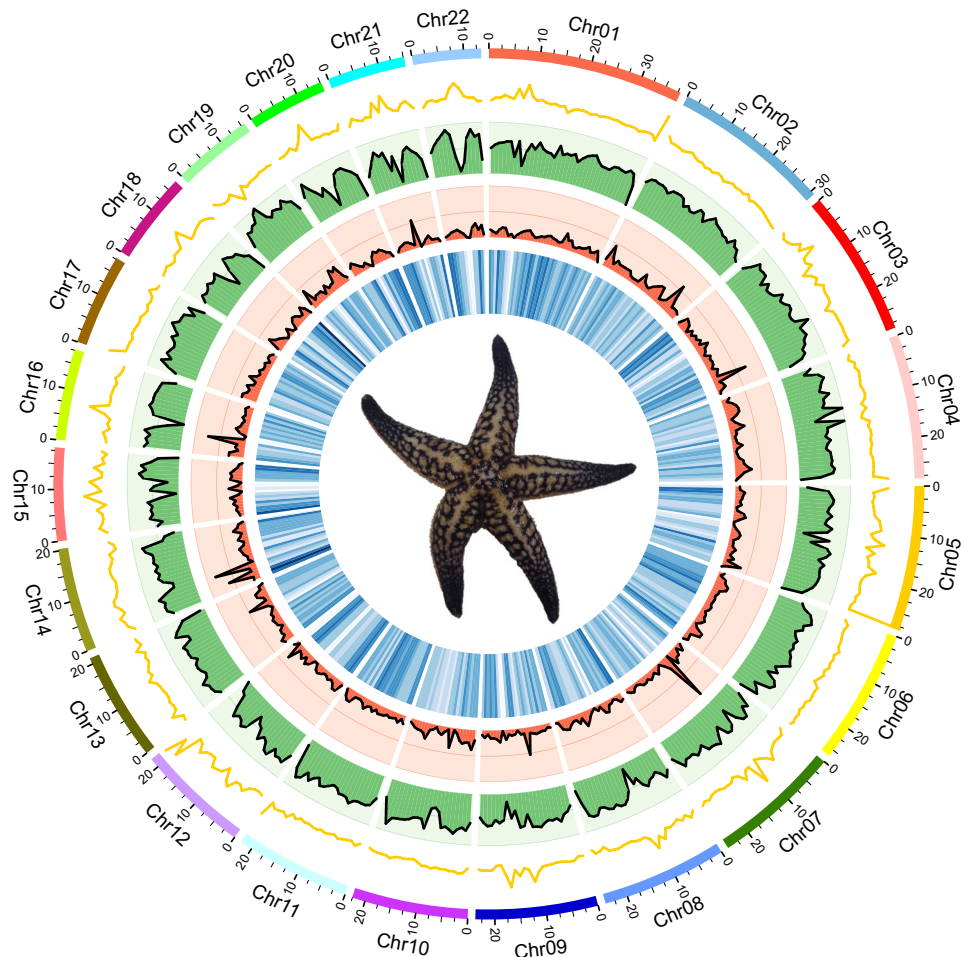
The chromosome conformation capture (Hi-C) technique was employed to assemble a chromosome-level genome. The fresh gonad was crosslinked using formaldehyde solution and digested with four-cutter restriction enzyme (DpnII). The ends of the restriction fragments were labeled with biotinylated nucleotides, and then the ligated DNA was sheared into fragments from 300 bp to 700 bp in length for Hi-C library construction. The resulting library was quantified with the Q-PCR method and sequenced with the DNBSEQ-T7 platform. After removing adapters and low-quality short reads, a total of 102.75 Gb ( $209.04\times$  coverage) of clean data was generated, with Q20 = 97.32% and Q30 = 92.33% (Table 2).

For transcriptome sequencing, total RNA of nine tissues from the same starfish was extracted and equally pooled for cDNA library construction. The resulting library was constructed by NEBNext<sup>®</sup> Ultra<sup>™</sup> RNA Library Prep Kit (NEB, USA) according to the manufacturer's instructions and sequenced on Illumina NovaSeq6000 system, finally generating 13.47 Gb clean data to help genome structure annotation.

**Genome assembly.** Based on PacBio HiFi reads, Hifiasm (v0.18.4)<sup>33</sup> was applied for *de novo* assembly of primary contigs with default parameters. Haplotypic and heterozygous duplication was removed using purge\_dups (v1.2.6)<sup>34</sup> with the parameter of cutoffs '-l 5 -m 18 -u 54'. A primary assembly was generated, consisting of 90 contigs spanning 491.50 Mb. N50 and the maximum contig length were 8.05 and 28.59 Mb, respectively (Table 3).

We further scaffolded the contigs using Hi-C sequencing data to obtain a high-quality chromosome-scale genome. Juicer (v1.6)<sup>35</sup> was applied for raw sequence data analysis and then 3D-DNA (v190716)<sup>36</sup> was used to anchor contigs into chromosomes. The assembly was further corrected manually according to the Hi-C heatmap using JuiceboxGUI (v1.11.08)<sup>37</sup>, a visualization system for Hi-C contact maps. The final genome consisted of 22 chromosomes with lengths ranging from 13.43 to 38.00 Mb, and the N50 was 23.75 Mb (Table 3, Fig. 1, Fig. 2). Previous karyotype analysis<sup>38</sup> of *A. amurensis* indicated that it had a diploid chromosome number of 44, which was consistent with our results.

**Annotation of repetitive elements.** The Extensive *de novo* TE Annotator (EDTA, v2.0.0)<sup>39</sup> and RepeatModeler (v2.0.3)<sup>40</sup> were utilized to build repetitive sequence libraries for *A. amurensis* genome. We combined these two libraries as a final comprehensive repeat library for repeat annotation. Then, RepeatMasker (v4.1.2)<sup>41</sup> was used to predict and classify repetitive elements of *A. amurensis* genome. Overall, sequences constituting 48.69% of the assembled genome were identified as repeats, of which the most abundant repetitive element was long terminal repeats (LTR, 19.63%), followed by DNA transposons (18.20%) (Table 4, Fig. 2).



**Fig. 2** Circos plot of genomic features in *A. amurensis* genome. The tracks from outside to inside indicate: (1) length of 22 chromosomes (Mb), (2) distribution of GC content with a window of 1 Mb, (3) distribution of repeat elements with a window of 1 Mb, (4) distribution of ncRNAs with a window of 1 Mb, and (5) distribution of protein-coding genes with a window of 1 Mb.

Type		Count	Length (bp)	% of Genome	
Dispersed repeats	DNA transposons	747,789	89,432,495	18.20	
	Retroelements	LTR	597,681	96,479,749	19.63
		LINE	2,928	1,233,195	0.25
		DIRS	416	206,153	0.04
		Penelope	2,072	791,234	0.16
Unclassified	118,869	45,699,346	9.30		
Tandem repeats	Simple repeats	82,117	4,897,016	1.00	
	Low complexity	11,639	592,691	0.12	
Total		1,563,511	239,331,879	48.69	

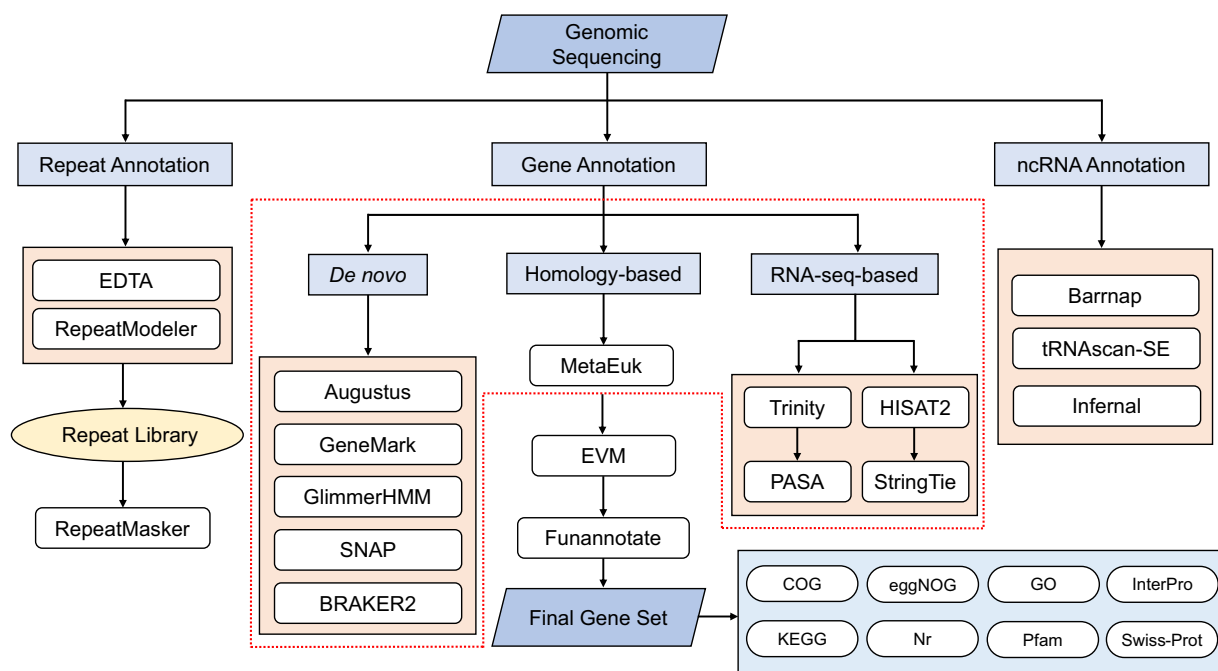
**Table 4.** Classification of repetitive sequences in *A. amurensis* genome.

**Noncoding RNA (ncRNA) annotation.** Ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs) were predicted by Barrnap (v0.9, <https://github.com/tseemann/barrnap>) and tRNAscan-SE (v2.0.11)<sup>42</sup> with default parameters, respectively. Based on an alignment with Rfam database (v14.8)<sup>43</sup>, Infernal (v1.1.4)<sup>44</sup> was used to annotate other ncRNAs, including small nuclear RNAs (snRNAs) and microRNAs (miRNAs). In total, we identified 37 miRNAs, 14,926 tRNAs, 415 rRNAs, and 202 snRNAs in *A. amurensis* genome (Table 5, Fig. 2).

**Gene prediction and functional annotation.** We used three approaches for predictions of gene structures, including *de novo*, homology-based, and RNA-seq-based prediction. Augustus (v3.4.0)<sup>45</sup>, GlimmerHMM

Type		Copy number	Average length(bp)	Total length(bp)	% of genome
miRNA		37	85.62	3,168	0.0006445
tRNA		14,926	72.28	1,078,878	0.2194907
rRNA	28 S	42	2,241.43	94,140	0.0191522
	18 S	22	1,813.00	39,886	0.0081145
	5.8 S	22	117.00	2,574	0.0005237
	5 S	329	99.19	32,633	0.0066390
snRNA	CD-box	54	98.24	5,305	0.0010793
	HACA-box	27	171.41	4,628	0.0009415
	scaRNA	1	94	94	0.0000191
	splicing	120	142.28	17,073	0.0034734

**Table 5.** Classification of ncRNAs in *A. amurensis* genome.

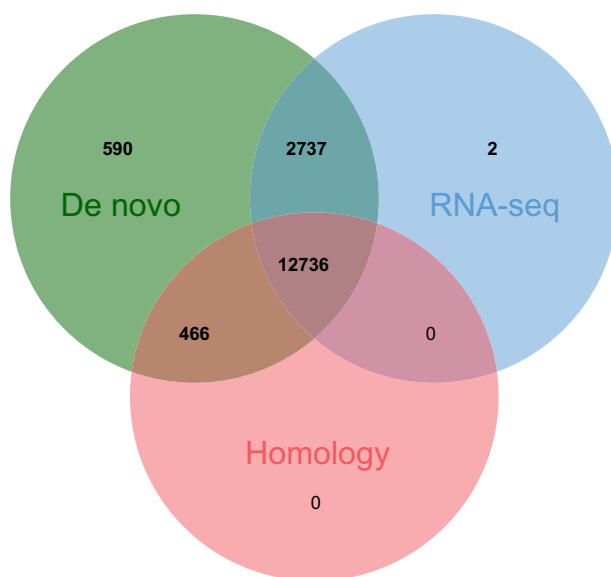


**Fig. 3** The general annotation pipeline of repetitive elements, ncRNAs, and protein-coding genes.

(v3.0.4)<sup>46</sup>, GeneMark (v4.69)<sup>47</sup>, SNAP (version 2006-07-28)<sup>48</sup>, and BRAKER2 (v2.1.6)<sup>49</sup> were utilized for *de novo* gene model prediction and they were performed with default parameters. For homology-based prediction, we downloaded protein sequences of the crown-of-thorns starfish *Acanthaster sp.* ([https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/001/949/145/GCF\\_001949145.1\\_OKI-Apl\\_1.0/](https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/001/949/145/GCF_001949145.1_OKI-Apl_1.0/)), sea urchin *Strongylocentrotus purpuratus* ([https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/002/235/GCF\\_000002235.5\\_Spur\\_5.0/](https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/002/235/GCF_000002235.5_Spur_5.0/)), and sea cucumber *Apostichopus japonicus* ([https://ftp.ncbi.nlm.nih.gov/genomes/genbank/invertebrate/Apostichopus\\_japonicus/latest\\_assembly\\_versions/GCA\\_002754855.1\\_ASM275485v1/](https://ftp.ncbi.nlm.nih.gov/genomes/genbank/invertebrate/Apostichopus_japonicus/latest_assembly_versions/GCA_002754855.1_ASM275485v1/)) from National Center for Biotechnology Information (NCBI) as references and used MetaEuk (version aa7ac2eb7334405ad57d50d78361e3dcd61bb27a)<sup>50</sup> with default parameters to predict gene structures. For RNA-seq-based prediction, we firstly mapped short RNA reads to reference genome using HISAT2 (v2.2.1)<sup>51</sup> with the parameter ‘-dta’ and then assembled transcripts using StringTie (v2.2.1)<sup>52</sup>. Meanwhile, the Program to Assemble Spliced Alignments (PASA, v2.4.1) pipeline (<https://github.com/PASApipeline/PASApipeline>) was used to identify possible coding regions based on *de novo* transcriptome assembled by Trinity (v2.14.0)<sup>53</sup> with default parameters. Then, EvidenceModeler (EVM, v1.1.1)<sup>54</sup> and Funannotate (v1.8.14) pipeline (<https://github.com/nextgenusfs/funannotate>) were applied for combining predicted results from three strategies and removal of low-quality gene annotations. Based on the RNA-seq data of *A. amurensis* from this study, adult stomach tissue<sup>55</sup>, and bipinnaria larval<sup>16</sup> from other studies, PASA (v2.4.1) was applied for the update of untranslated regions (UTRs). The general annotation pipeline applied in the present study was shown in Fig. 3. As a result, a total of 16,531 protein-coding genes were predicted and the average gene length was 17,803.19 bp, with an average coding sequence (CDS) length of 1,885.87 bp and average exon number of 10.07 (Table 6). Among them, 12,736 (77.04%) genes were supported by evidence from all three strategies

	Gene set	Gene Number	Gene length (bp)	CDS length (bp)	Average intron length (bp)	Average exon length (bp)	Exon per gene
De novo	Augustus	20,742	12,558.23	1,743.97	1,466.01	208.19	8.38
	GlimmerHMM	59,965	7,146.75	847.98	2,106.12	197.82	4.29
	GeneMark	22,312	9,008.46	1,587.86	1,172.18	216.61	7.33
	SNAP	25,901	26,508.15	2,030.64	2,200.07	167.47	12.13
	BRAKER2	24,353	10,549.42	1,702.42	1,848.07	206.99	8.22
RNA-seq	PASA	14,250	14,259.92	1,315.87	1,970.86	298.97	7.15
	HISAT2 & StringTie	15,047	21,787.79	1,922.51	2,005.33	383.69	12.26
Homology	<i>A. planci</i>	10,063	11,124.46	1,309.15	2,680.73	282.37	4.66
	<i>A. japonicus</i>	6,200	5,562.53	956.09	2,715.86	355.90	2.67
	<i>S. purpuratus</i>	7,265	6,095.00	1,106.24	2,621.27	382.42	2.90
Final		16,531	17,803.19	1,885.87	1,743.54	283.00	10.07

**Table 6.** Statistical results of the gene structure annotation in *A. amurensis* genome.

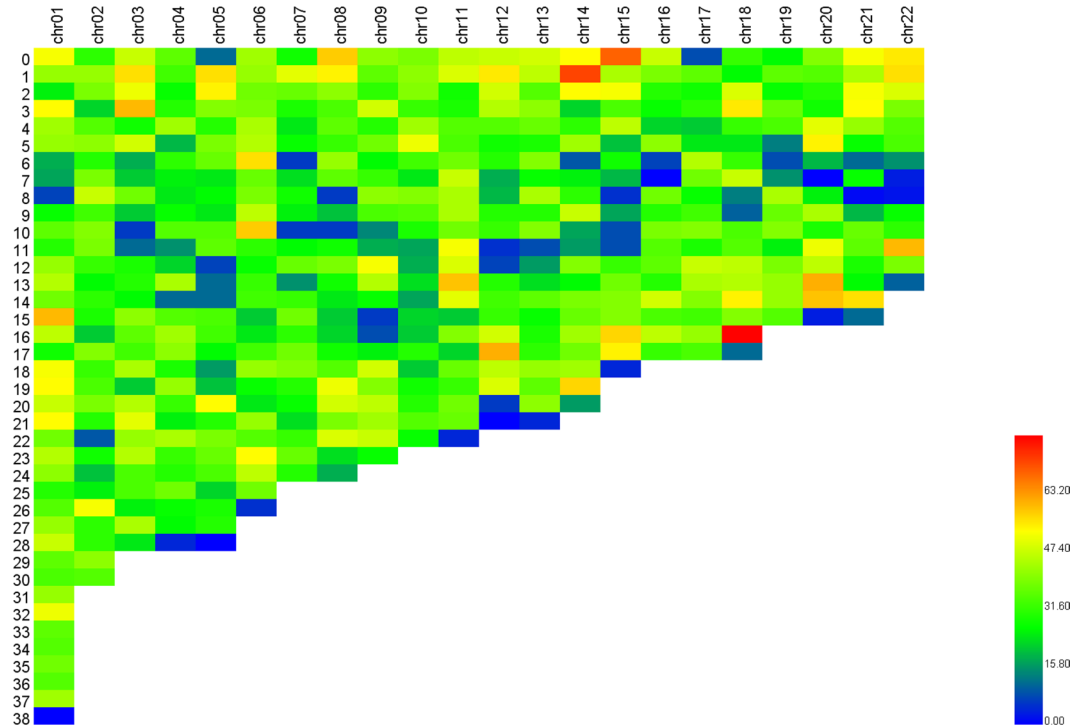


**Fig. 4** Venn diagram of gene structure prediction from *de novo*, homology-based and RNA-seq-based strategies.

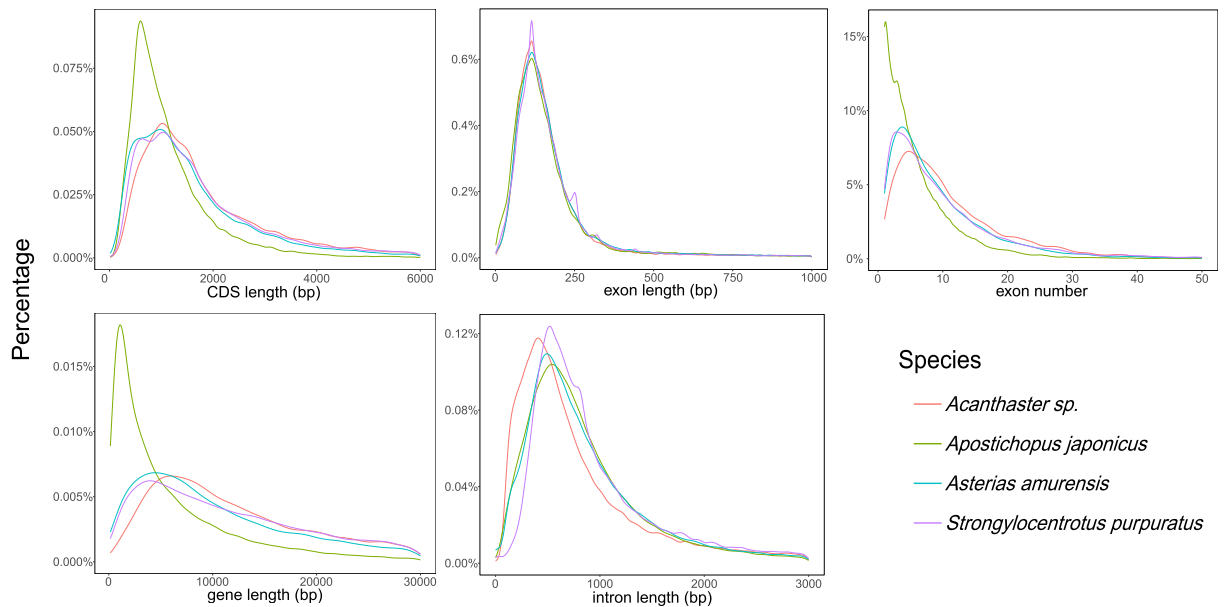
(Fig. 4). We also counted the density of genes on different chromosomes with a window of 1 Mb in length (Fig. 5) and simply compared gene length, CDS length, exon length, intron length and exon number per gene of *A. amurensis* and other species used in homology-based predictions (Fig. 6). The 1 Mb region with the largest number of annotated genes were from the end of chromosome 18 (Fig. 5).

Functional annotations were accomplished using Funannotate pipeline, based on databases including Clusters of Orthologous Groups of Proteins (COG)<sup>56</sup>, eggNOG<sup>57</sup>, Gene Ontology (GO)<sup>58</sup>, Interpro<sup>59</sup>, Kyoto Encyclopedia of Genes and Genomes (KEGG)<sup>60</sup>, NCBI non-redundant protein (Nr), Pfam<sup>61</sup>, and Swiss-Prot<sup>62</sup>. The results showed that 15,643 protein sequences (94.63%) were annotated with at least one public database (Table 7, Fig. 7).

**Comparative genomic analysis.** The longest protein sequences of *A. amurensis* and other five asteroid species including *Acanthaster sp.*<sup>63</sup>, *Asterias rubens*<sup>64</sup>, *Patiria miniata*<sup>65</sup>, *Plazaster borealis*<sup>66</sup>, and *Zoroaster cf. ophiactis*<sup>67</sup> were utilized to identify orthologous groups using OrthoFinder (v2.5.5)<sup>68</sup> with the parameters ‘-S diamond’, and the sea urchin *Lytechinus variegatus*<sup>69</sup> was selected as an outgroup. A total of 5,315 single-copy orthogroups were obtained for subsequent phylogenetic analysis. Based on multiple sequence alignments of the single-copy orthogroups using MAFFT (v7.520)<sup>70</sup>, IQ-TREE (v2.2.3)<sup>71</sup> was applied for construction of the species trees with the parameters ‘-m MFP -bb 1000’ and the best model of GTR + F + I + R4. Predictably, *A. amurensis* was most closely related to *A. rubens* and *P. borealis* from the family Asteroidea (Fig. 8). Then, divergence times were estimated using MCMCTREE in PAML (v4.9i)<sup>72</sup> based on the divergence time (*A. amurensis* vs *L. variegatus*: 461.1.5–600.0 million years ago) extracted from TIMETREE (<http://www.timetree.org/>). The expansion and contraction of gene families were analyzed by Computational Analysis of gene Family Evolution (CAFE, v5.0.0)<sup>73</sup>



**Fig. 5** Number of genes on 22 chromosomes with a window of 1 Mb. The scale bar represents the density of genes.



**Fig. 6** Comparisons of CDS length, exon length, exon number per gene, gene length and intron length among *A. amurensis* and other relative species.

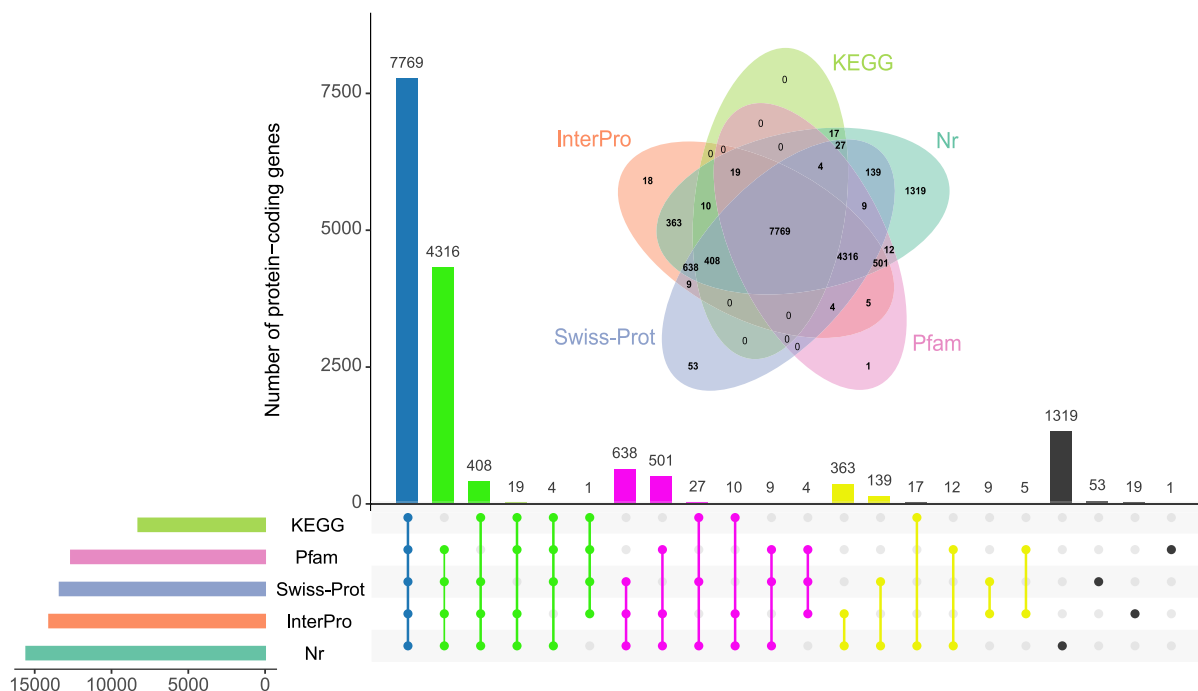
with a p-value of 0.05. The results revealed that 197 and 482 gene families were expanded and contracted in *A. amurensis*, respectively (Fig. 8).

### Data Records

The PacBio, BGI, RNA-seq, and Hi-C sequencing data have been deposited in the NCBI Sequence Read Archive (SRA) database under the accession numbers of SRR24902114<sup>74</sup>, SRR24831139<sup>75</sup>, SRR24871501<sup>76</sup>, and SRR24835318<sup>77</sup>. The final chromosome assembly has been deposited in GenBank with assembly accession number GCA\_032118995.1<sup>78</sup>. The genome annotation files are available in the Figshare database<sup>79</sup>.

Database	Number	Percent(%)
Total	16,531	—
COG	13,657	82.61
EggNOG	14,245	86.17
GO	10,373	62.75
InterPro	14,060	85.05
KEGG	8,254	49.93
Pfam	12,640	76.46
Swiss-Prot	13,376	80.91
Nr	15,551	94.07
Annotated	156,43	94.63
Unannotated	888	5.37

**Table 7.** Summary of the functional gene annotation in *A. amurensis* genome.



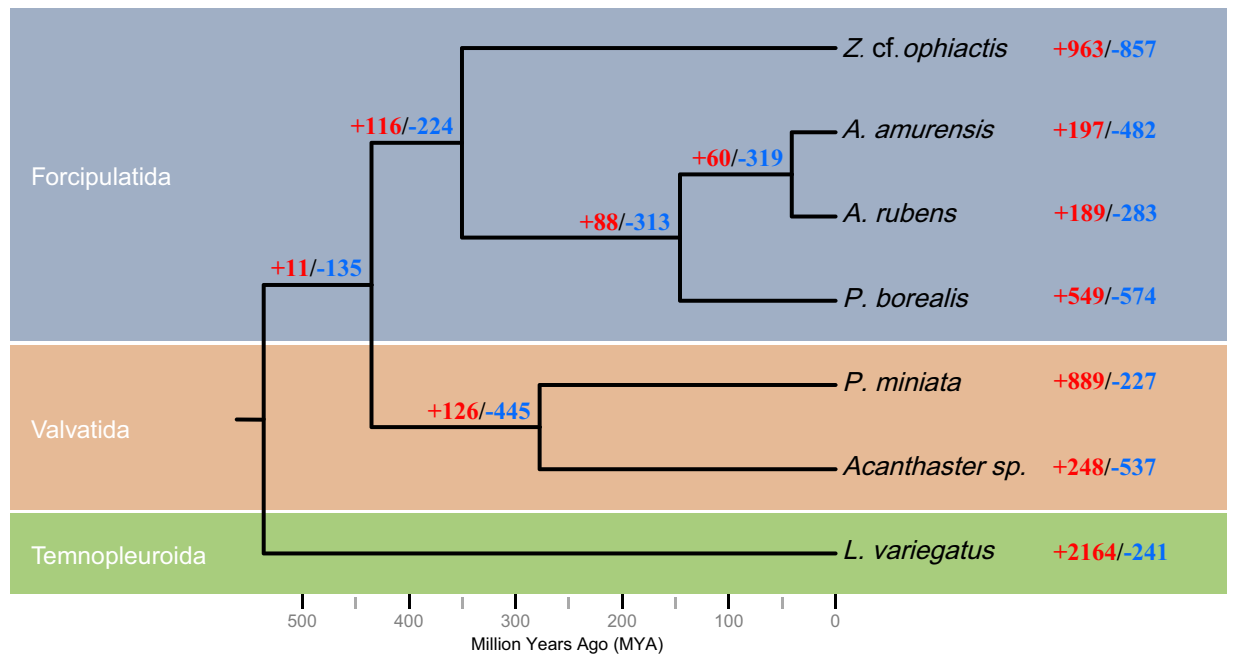
**Fig. 7** Upset plot and Venn diagram of functional annotation for protein-coding genes based on different databases, including InterPro, KEGG, Nr, Pfam, and Swiss-Prot.

### Technical Validation

**Nucleic acid quality.** The concentration and quality of DNA were evaluated using Nanodrop 2000 spectrophotometer (Thermo Fisher Scientific, USA) and agarose gel electrophoresis, respectively. RNA integrity was assessed using Agilent 2100 Bioanalyzer (Agilent Technologies, USA).

**Genome assembly and annotation quality evaluation.** The quality of the final chromosome-level genome assembly was assessed using four methods as follows. Firstly, we mapped clean PE150 reads from whole genome sequencing to *A. amurensis* genome using BWA-MEM (v0.7.17)<sup>80</sup> and calculated the mapping rate using samtools (v1.9)<sup>81</sup>, resulting in a genome coverage rate of 99.95% and a mapping rate of 99.61%. Secondly, the results of Benchmarking Universal Single-Copy Orthologs (BUSCO, v5.2.2)<sup>82</sup> analysis based on 954 genes of metazoa\_odb10 database indicated that 951 (99.69%) core metazoan genes were detected in *A. amurensis* genome, consisting of 943 (98.85%) complete and 8 (0.84%) fragmented genes (Table 8). Thirdly, the Core Eukaryotic Genes Mapping Approach (CEGMA, v2.5)<sup>83</sup> based on 248 core eukaryotic genes showed that 236 (95.16%) genes were identified in the final genome assembly. Finally, meryl (v1.3)<sup>84</sup> was used to generate k-mer counts based on paired-end reads generated by whole genome sequencing, and Merqury (v1.3)<sup>84</sup> was utilized to estimate the consensus quality value (QV) of *A. amurensis* genome, resulting in a QV of 48.51. The results from the four methods above revealed the high accuracy and completeness of the final genome assembly.





**Fig. 8** Phylogenetic and gene family evolution analysis between *A. amurensis* and the other five asteroid species. The sea urchin *L. variegatus* was selected as an outgroup. All species were colored according to different orders. The scale below represents the divergence time. The number of expanded (+red) and contracted (-blue) gene families were shown alongside the species.

Type	Percentage
Complete BUSCOs (C)	98.85% (943)
Complete and single-copy BUSCOs (S)	97.80% (933)
Complete and duplicated BUSCOs (D)	1.05% (10)
Fragmented BUSCOs (F)	0.84% (8)
Missing BUSCOs (M)	0.31% (3)
Total	100% (954)

**Table 8.** BUSCO evaluation of gene annotation in *A. amurensis* genome.

### Code availability

No custom code was utilized in this study. Data processing was performed by relevant pipelines and software according to the manual and protocols and the version as well as useful parameters have been described in the Methods section. The default parameters as developers suggested were used in those pipelines and software of which parameters were not specifically mentioned in this work.

Received: 6 July 2023; Accepted: 25 October 2023;

Published online: 04 November 2023

### References

- Fukuyama, A. K. & Oliver, J. S. Sea star and walrus predation on bivalves in Norton Sound, Bering Sea, Alaska. *Ophelia* **24**, 17–36 (1985).
- Li, B. *et al.* Size distribution of individuals in the population of *Asterias amurensis* (Echinodermata: Asteroidea) and its reproductive cycle in China. *Acta Oceanol. Sin.* **37**, 96–103 (2018).
- Ward, R. D. & Andrew, J. Population genetics of the northern Pacific seastar *Asterias amurensis* (Echinodermata: Asteroidea): allozyme differentiation among Japanese, Russian, and recently introduced Tasmanian populations. *Mar. Biol.* **124**, 99–109 (1995).
- Paik, S. G., Park, H. S., Yi, S. K. & Yun, S. G. Developmental duration and morphology of the sea star *Asterias amurensis*, in Tongyeong, Korea. *Ocean Sci. J.* **40**, 65–70 (2005).
- Kashenko, S. D. Responses of embryos and larvae of the starfish *Asterias amurensis* to changes in temperature and salinity. *Russ. J. Mar. Biol.* **31**, 294–302 (2005).
- Reich, A., Dunn, C., Akasaka, K. & Wessel, G. Phylogenomic analyses of Echinodermata support the sister groups of Asterozoa and Echinozoa. *PLoS One* **10**, e0119627 (2015).
- Dupont, S. & Thorndyke, M. Bridging the regeneration gap: insights from echinoderm models. *Nat. Rev. Genet.* **8**, 320–320 (2007).
- Medina-Feliciano, J. G. & Garcia-Ararras, J. E. Regeneration in echinoderms: molecular advancements. *Front. Cell Dev. Biol.* **9**, 768641 (2021).
- Byrne, M., Morrice, M. G. & Wolf, B. Introduction of the northern Pacific asteroid *Asterias amurensis* to Tasmania: reproduction and current distribution. *Mar. Biol.* **127**, 673–685 (1997).

10. Kashenko, S. D. Development of the starfish *Asterias amurensis* under laboratory conditions. *Russ. J. Mar. Biol.* **31**, 36–42 (2005).
11. Qu, P. *et al.* Trophic structure of common marine species in the Bohai Strait, North China Sea, based on carbon and nitrogen stable isotope ratios. *Ecol. Indic.* **66**, 405–415 (2016).
12. Hutson, K. S., Ross, D. J., Day, R. W. & Ahern, J. J. Australian scallops do not recognise the introduced predatory seastar *Asterias amurensis*. *Mar. Ecol. Prog. Ser.* **298**, 305–309 (2005).
13. Ross, D. J., Johnson, C. R. & Hewitt, C. L. Impact of introduced seastars *Asterias amurensis* on survivorship of juvenile commercial bivalves *Fulvia tenuicostata*. *Mar. Ecol. Prog. Ser.* **241**, 99–112 (2002).
14. Nishimura, H., Miyoshi, K. & Chiba, S. Predatory behavior of the sea stars *Asterias amurensis* and *Distolasterias nipon* on the Japanese scallop, *Mizuhopecten yessoensis*. *Plankton Benthos Res.* **14**, 1–7 (2019).
15. Parry, G. D. & Hirst, A. J. Decadal decline in demersal fish biomass coincident with a prolonged drought and the introduction of an exotic starfish. *Mar. Ecol. Prog. Ser.* **544**, 37–52 (2016).
16. Richardson, M. F. & Sherman, C. D. De novo assembly and characterization of the invasive northern Pacific seastar transcriptome. *PLoS One* **10**, e0142003 (2015).
17. Dunstan, P. K. & Bax, N. J. How far can marine species go? Influence of population biology and larval movement on future range limits. *Mar. Ecol. Prog. Ser.* **344**, 15–28 (2007).
18. Ling, S. D., Johnson, C. R., Mundy, C. N., Morris, A. & Ross, D. J. Hotspots of exotic free-spawning sex: man-made environment facilitates success of an invasive seastar. *J. Appl. Ecol.* **49**, 733–741 (2012).
19. Ross, D. J., Johnson, C. R. & Hewitt, C. L. Abundance of the introduced seastar, *Asterias amurensis*, and spatial variability in soft sediment assemblages in SE Tasmania: clear correlations but complex interpretation. *Estuarine, Coastal Shelf Sci.* **67**, 695–707 (2006).
20. Hayes, K. R. & Sliwa, C. Identifying potential marine pests—a deductive approach applied to Australia. *Mar. Pollut. Bull.* **46**, 91–98 (2003).
21. Ellis, M. R. *et al.* Detecting marine pests using environmental DNA and biophysical models. *Sci. Total Environ.* **816**, 151666 (2022).
22. Richardson, M. F., Sherman, C. D., Lee, R. S., Bott, N. J. & Hirst, A. J. Multiple dispersal vectors drive range expansion in an invasive marine species. *Mol. Ecol.* **25**, 5001–5014 (2016).
23. Byrne, M., Gall, M., Wolfe, K. & Agüera, A. From pole to pole: the potential for the Arctic seastar *Asterias amurensis* to invade a warming Southern Ocean. *Glob. Chang. Biol.* **22**, 3874–3887 (2016).
24. Bock, D. G. *et al.* What we still don't know about invasion genetics. *Mol. Ecol.* **24**, 2277–2297 (2015).
25. Ross, D. J., Johnson, C. R. & Hewitt, C. L. Assessing the ecological impacts of an introduced seastar: the importance of multiple methods. *Biol. Invasions* **5**, 3–21 (2003).
26. Li, L., Yu, Y., Wu, W. & Wang, P. Extraction, characterization and osteogenic activity of a type I collagen from starfish (*Asterias amurensis*). *Mar. Drugs* **21**, 274 (2023).
27. Witman, J. D., Genovese, S. J., Bruno, J. F., McLaughlin, J. W. & Pavlin, B. I. Massive prey recruitment and the control of rocky subtidal communities on large spatial scales. *Ecol. Monogr.* **73**, 441–462 (2003).
28. Smith, K. F. *et al.* Application of a sandwich hybridisation assay for rapid detection of the northern Pacific seastar, *Asterias amurensis*. *N. Z. J. Mar. Freshwater Res.* **45**, 145–152 (2011).
29. Pochon, X., Bott, N. J., Smith, K. F. & Wood, S. A. Evaluating detection limits of next-generation sequencing for the surveillance and monitoring of international marine pests. *PLoS One* **8**, e73935 (2013).
30. Deagle, B. E., Bax, N., Hewitt, C. L. & Patil, J. G. Development and evaluation of a PCR-based test for detection of *Asterias* (Echinodermata: Asteroidea) larvae in Australian plankton samples from ballast water. *Mar. Freshwater Res.* **54**, 709–719 (2003).
31. Hall, M. R. *et al.* The crown-of-thorns starfish genome as a guide for biocontrol of this coral reef pest. *Nature* **544**, 231–234 (2017).
32. Xu, C. *et al.* Chromosome level genome assembly of oriental armyworm *Mythimna separata*. *Sci. Data* **10**, 597 (2023).
33. 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 (2021).
34. Guan, D. *et al.* Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).
35. Durand, N. C. *et al.* Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
36. Dudchenko, O. *et al.* De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
37. Durand, N. C. *et al.* Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101 (2016).
38. Saotome, K. & Komatsu, M. Chromosomes of Japanese starfishes. *Zool. Sci.* **19**, 1095–1103 (2002).
39. Ou, S. *et al.* Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).
40. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA* **117**, 9451–9457 (2020).
41. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinf.* Chapter 4, 10.1–10.14 (2009).
42. Chan, P. P., Lin, B. Y., Mak, A. J. & Lowe, T. M. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res.* **49**, 9077–9096 (2021).
43. Kalvari, I. *et al.* Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res.* **49**, D192–D200 (2021).
44. Nawrocki, E. P., Kolbe, D. L. & Eddy, S. R. Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25**, 1335–1337 (2009).
45. Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
46. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
47. Besemer, J. & Borodovsky, M. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res.* **33**, W451–W454 (2005).
48. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
49. Bruna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics Bioinf.* **3**, lqaa108 (2021).
50. Levy Karin, E., Mirdita, M. & Soding, J. MetaEuk—sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. *Microbiome* **8**, 48 (2020).
51. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
52. Kovaka, S. *et al.* Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278 (2019).
53. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
54. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVIDENCEModeler and the program to assemble spliced alignments. *Genome Biol.* **9**, R7 (2008).
55. *NCBI Sequence Read Archive* <https://identifiers.org/ncbi/insdc.sra:SRR26081154> (2023).
56. Galperin, M. Y. *et al.* COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res.* **49**, D274–D281 (2021).
57. Cantalapiedra, C. P., Hernandez-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).

58. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
59. Blum, M. *et al.* The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* **49**, D344–D354 (2021).
60. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–D462 (2016).
61. Mistry, J. *et al.* Pfam: the protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
62. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45–48 (2000).
63. Baughman, K. W. *et al.* *Acanthaster planci*, whole genome shotgun sequencing project. *GenBank* <https://identifiers.org/ncbi/insdc:BDGF01000000> (2016).
64. Wellcome Sanger Institute. *Asterias rubens*, whole genome shotgun sequencing project. *GenBank* <https://identifiers.org/ncbi/insdc:CABPRM03000000> (2019).
65. Ku, C. J., Cary, G. A. & Hinman, V. F. *Patiria miniata* isolate m\_02\_andy, whole genome shotgun sequencing project. *GenBank* <https://identifiers.org/ncbi/insdc:JADOBP01000000> (2020).
66. Lee, Y. *et al.* Chromosome-level genome assembly of *Plazaster borealis* sheds light on the morphogenesis of multiarmed starfish and its regenerative capacity. *GigaScience* **11**, giac063 (2022).
67. Liu, J., Zhou, Y., Pu, Y. & Zhang, H. A chromosome-level genome assembly of a deep-sea starfish (*Zoroaster cf. ophiactis*). *Sci. Data* **10**, 506 (2023).
68. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015).
69. Davidson, P. L. *et al.* Chromosomal-level genome assembly of the sea urchin *Lytechinus variegatus* substantially improves functional genomic analyses. *Genome Biol. Evol.* **12**, 1080–1086 (2020).
70. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
71. Minh, B. Q. *et al.* IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
72. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
73. Mendes, F. K., Vanderpool, D., Fulton, B. & Hahn, M. W. CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics* **36**, 5516–5518 (2020).
74. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR24902114> (2023).
75. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR24831139> (2023).
76. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR24871501> (2023).
77. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR24835318> (2023).
78. Wang, Y. L. *et al.* Chromosome-level genome assembly of northern Pacific seastar *Asterias amurensis*. *GenBank* [https://www.ncbi.nlm.nih.gov/assembly/GCA\\_032118995.1](https://www.ncbi.nlm.nih.gov/assembly/GCA_032118995.1) (2023).
79. Wang, Y. L. *et al.* Chromosome-level genome assembly of *Asterias amurensis*. *figshare*. <https://doi.org/10.6084/m9.figshare.23538585.v2> (2023).
80. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
81. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
82. Manni, M., Berkeley, M. R., Seppely, M., Simao, 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 (2021).
83. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067 (2007).
84. 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 (2020).

## Acknowledgements

We thank Prof. Scott F Cummins for his help in manuscript polishing. This research was supported by National Natural Science Foundation of China [grant numbers 31972767 and 42276103].

## Author contributions

M.Y.C. conceived the study. Y.L.W. prepared the nucleic acid samples. Y.L.L. performed the data analysis. Y.L.W., Y.X.W., Y.J.Y. and G.N. visualized the results and wrote the manuscript. All authors reviewed and approved the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to Y.L. or M.C.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

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



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023