



OPEN

DATA DESCRIPTOR

# Chromosome-level reference genome and annotation of the Arctic fish *Anisarchus medius*

Ruoyu Liu<sup>1</sup>, Ziyu Meng<sup>1</sup>, Yinan Mu<sup>1</sup>, Ran Zhang<sup>2</sup>, Hanhui Ma<sup>1</sup>, Jingjing Hu<sup>1</sup>, Yanan Wang<sup>1</sup>, Yuxin Shi<sup>1</sup>, Yanan Li<sup>3</sup>, Chaofeng Wang<sup>1</sup>, Weini Zhang<sup>1</sup>, Longshan Lin<sup>2</sup>✉, Ping Zheng<sup>1</sup>✉ & Xinhua Chen<sup>1</sup>✉

*Anisarchus medius* (Reinhardt, 1837) is a widely distributed Arctic fish, serving as an indicator of climate change impacts on coastal Arctic ecosystems. This study presents a chromosome-level genome assembly for *A. medius* using PacBio sequencing and Hi-C technology. The PacBio assembly totaled 739.07 Mb across 697 contigs, with a Contig N50 of 10.004 Mb. Hi-C mapping yielded 23 chromosomes, with a successful mapping rate of 90.53% and a Scaffold N50 of 30.20 Mb. Genome BUSCO integrity was 97.05%. Repetitive sequences accounted for 240.83 Mb (32.58%) of the genome. Non-coding RNA annotations included 4,928 rRNAs, 9,663 tRNAs, 347 snRNAs, and 21 snoRNAs. A total of 30,345 protein-coding genes were identified, encoding 46,603 proteins, with a BUSCO completeness of 94.98%. Molecular pathway related to the endocrine system, carbohydrate metabolism, folding, sorting, and degradation, signal transduction, and transport and catabolism contribute to *A. medius* adaptation to extreme Arctic environments. This high-quality genome provides valuable genetic resources for understanding Arctic adaptations and supporting polar ecological conservation and management.

## Background & Summary

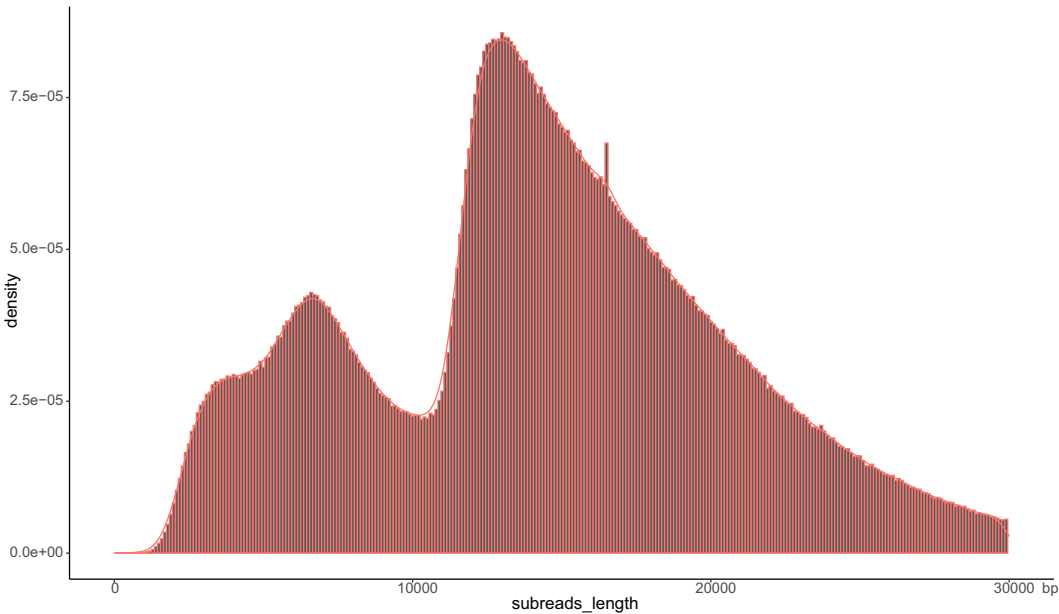
*Anisarchus medius*, commonly known as stout eelblenny, is a ray-fin fish that belongs to the eel pricklebaks, taxonomically classified under Order Perciformes, Suborder Cottioidei, Infraorder Zoarcales, and Family Stichaeidae (NCBI: txid980385, Fig. 1). *A. medius* is a common seen species in the Arctic Ocean, with a practically circumpolar distribution and particular abundance in the Pacific Arctic<sup>1</sup>. Thriving in muddy substrates at depths of 30–100 m<sup>1</sup> and favoring water temperatures below 0 °C<sup>2,3</sup>, this benthic fish primarily feeds on polychaetes and other small benthic organisms<sup>1,4</sup>, acting as a critical intermediary in transferring energy from lower benthic trophic levels to higher predators<sup>4</sup>. By maintaining ecosystem stability, *A. medius* plays an essential role in Arctic marine food webs<sup>1,4</sup>. Given the Arctic ecosystem is highly sensitive to climate change, with rising temperatures and shifting ice dynamics threatening its biodiversity<sup>5–7</sup>, *A. medius* has emerged as an important indicator species for assessing the climate change on population dynamics within the coastal Arctic ecosystems<sup>2</sup>. Deciphering the high-quality genome of the ecosystem indicator species like *A. medius* is critical for informing future conservation and management strategies in the face of a rapidly warming climate.

Despite its ecological importance, research on *A. medius* remains limited and predominantly focused on fundamental aspects, including morphological classification, growth and development, geographical distribution, phylogeny, nutrient composition, and ecological and climatic studies. Morphological studies address characteristics such as body size, age description<sup>1,8,9</sup>, and skeletal development during ontogenesis<sup>10</sup>. Growth and development studies have explored various life stages, including egg, larval, post-larval, and juvenile phases<sup>1,10,11</sup>. Geographical distribution surveys<sup>1,9</sup> and phylogenetic research based on mitochondrial genomes and conserved

<sup>1</sup>State Key Laboratory of Mariculture Breeding, Key Laboratory of Marine Biotechnology of Fujian Province, Institute of Oceanology, College of Marine Sciences, Haixia Institute of Science and Technology, Fujian Agriculture and Forestry University, Fuzhou, 350002, China. <sup>2</sup>Laboratory of Marine Biodiversity Research, Third Institute of Oceanography, Ministry of Natural Resources, Xiamen, 361005, China. <sup>3</sup>Guangzhou Key Laboratory of Aquatic Animal Diseases and Waterfowl Breeding, College of Animal Sciences and Technology, Zhongkai University of Agriculture and Engineering, Guangzhou, 510222, China. ✉e-mail: [linlongshan@tio.org.cn](mailto:linlongshan@tio.org.cn); [zhengping13@mailsucas.ac.cn](mailto:zhengping13@mailsucas.ac.cn); [chenxinhua@tio.org.cn](mailto:chenxinhua@tio.org.cn)



**Fig. 1** The image of Stout Eelblenny (*Anisarchus medius*) collected in the Chukchi Sea (−169.02 °W; 67.68°N).



**Fig. 2** The length distribution of PacBio HiFi reads for *A. medius*.

Sample name	Read number	Total length (bp)	N50 (bp)	N90 (bp)	Mean (bp)	Median (bp)	Max. (bp)	Min. (bp)
HIFI_data	2,444,493	35,749,145,145	16,902	9,921	14,624	14,366	72,732	76

**Table 1.** The summary of HIFI-Subread statistic of *A. medius*.

sequences<sup>2,12</sup> have also been conducted. Physiological studies are scarce and mainly focus on the aspect such as fatty acid composition<sup>13,14</sup>. Key biological questions, such as the reproductive habits and life history of *A. medius*, remain unanswered. Ecological and climatic studies of *A. medius* are relatively more abundant, primarily addressing temperature-related phenomena, such as osmotic and ionic regulation at varying temperatures<sup>15</sup> and population migrations driven by global warming<sup>16–19</sup>. Genomic resources for *A. medius* are currently limited to one mitochondrial genome<sup>2</sup> and two genomic Illumina sequencing datasets<sup>20</sup>. Critical gaps in understanding the origin, evolution, and adaptive mechanisms of *A. medius* to the Arctic environment, as well as its conservation and management, highlight the urgent need for high-quality genome sequencing.

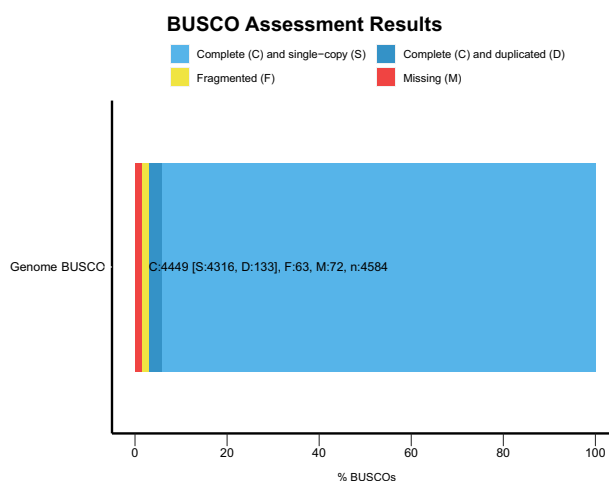
This study combines PacBio long-read sequencing with high-throughput chromosome conformation capture (Hi-C) technology to construct a high-quality, chromosome-level reference genome for *A. medius*. Comprehensive genomic annotation includes repetitive sequences, non-coding elements, and protein-coding genes. Functional annotation, including Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis, further enriches our understanding of the *A. medius* genome. These findings provide valuable insights into the adaptive evolution of Arctic organisms and establish a foundation for polar ecological conservation and management efforts.

Methods

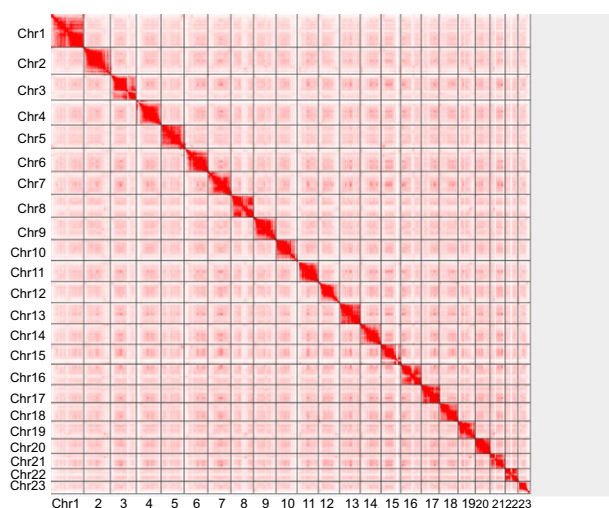
**Ethics statement.** All animal handling and experimental procedures were approved by the Fujian Agriculture and Forestry University Ethics Committee (approval number: PZCASFAFU24069).

Seq. type	All	Chr.	Contig.
Sep. number	390	23	697
Total length (bp)	739,230,572	669,270,049	739,077,072
N50 (bp)	30,207,471	30,887,387	10,004,925
L50 (bp)	10	9	22
N90 (bp)	18,504,769	23,751,151	591,375
L90 (bp)	22	19	155
Mean (bp)	1,895,463	29,098,697	1,060,368
Median (bp)	74,557	30,207,471	118,921
Max. (bp)	46,974,980	46,974,980	27,142,426
Min. (bp)	9,758	17,251,295	9,758
GC content (%)	42.36	42.26	42.37

**Table 2.** The assembly statistic of *A. medius*.



**Fig. 3** BUSCO assessment results for *A. medius* genome. Indicator: C, indicates complete BUSCOs; S, single-copy BUSCOs; D, complete and duplicated BUSCOs; F, fragmented BUSCOs; and M, missing BUSCOs; n, total number of BUSCO groups analyzed.



**Fig. 4** *A. medius* genome contig contact matrix was constructed using Hi-C data, with Chr1-23 representing the 23 chromosomes. The intensity of the red coloration indicates the contact density, where darker shades correspond to higher interaction frequencies.

Chr	Chr size (bp)	Contig number	Contig size (bp)	GC content (%)
Chr1	46,974,980	21	46,964,980	42.53
Chr2	38,175,894	15	38,168,894	41.72
Chr3	35,296,513	9	35,292,513	42.31
Chr4	34,856,473	20	34,846,973	42.24
Chr5	33,287,400	10	33,282,900	41.84
Chr6	32,028,287	19	32,019,287	42.14
Chr7	31,962,062	11	31,957,062	41.82
Chr8	31,274,129	17	31,266,129	41.97
Chr9	30,887,387	14	30,880,887	42.02
Chr10	30,207,471	19	30,198,471	42.36
Chr11	29,459,091	7	29,456,091	42.11
Chr12	29,152,110	22	29,141,610	42.48
Chr13	29,019,092	15	29,012,092	42.27
Chr14	28,372,503	6	28,370,003	42.04
Chr15	28,361,610	8	28,358,110	43.09
Chr16	28,092,748	10	28,088,248	42.24
Chr17	25,515,609	6	25,513,109	41.84
Chr18	25,127,810	3	25,126,810	42.19
Chr19	23,751,151	18	23,742,651	42.42
Chr20	21,254,401	6	21,251,901	42.02
Chr21	20,457,264	4	20,455,764	42.57
Chr22	18,504,769	7	18,501,769	43.61
Chr23	17,251,295	18	17,242,795	43.30
All	669,270,049	285	669,139,049	42.26

**Table 3.** The summary of *A. medius* chromosome lengths statistics.

**Sample collection and sequencing.** During the 2014 Chinese Arctic Research Expedition, A Stout Eelblenny (*A. medius*) was sampled aboard by the icebreaker R/V Xuelong (Snow Dragon) in the Chukchi Sea (−169.02 °W; 67.68°N) through trawling. The arctic fish sample was rapidly frozen with liquid nitrogen on the deck and then transferred to a −80 °C freezer for storage<sup>21–25</sup>. Upon returning to the laboratory, the dorsal muscle tissue of the fish was collected for genomic sequencing and analysis<sup>21</sup>. High-quality genomic DNA was extracted from the dorsal muscle tissue using QIAGEN Genomic-tip 100/G kit (Qiagen, Hilden, Germany)<sup>22</sup>. The quality and quantity of the extracted DNA were assessed using a NanoDrop 2000 spectrophotometer (NanoDrop Technologies, DE, USA), measuring the OD260/280 and OD260/230 ratios<sup>22,24</sup>. The concentration of the genomic DNA was accurately measured using a Qubit 3.0 Fluorometer (Life Technologies, CA, USA). The integrity of the genomic DNA was verified through 0.8% agarose gel electrophoresis<sup>21,22,24</sup>.

The SMRTbell library was prepared using the SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences)<sup>21,26</sup>. In brief, 15 µg of genomic DNA was subjected to an initial enzymatic reaction to remove single-stranded overhangs, followed by treatment with repair enzymes to fix any potential damage to the DNA backbone. Once the DNA damage was repaired, the ends of the double-stranded fragments were polished, and an A-overhang was added. The fragments were then ligated to T-overhang SMRTbell adapters at 20 °C for 60 min. After ligation, the SMRTbell library was purified using 1X AMPure PB beads. The library's size distribution and concentration were evaluated using a FEMTO Pulse automated pulsed-field capillary electrophoresis instrument (Agilent Technologies, DE, USA) and a Qubit 3.0 Fluorometer (Life Technologies, CA, USA). Once the library characterization was complete, 3 µg of the library underwent size selection using the BluePippin system (Sage Science, MA, USA), which removed SMRTbells smaller than 15 kb. After size selection, the library was again purified with 1X AMPure PB beads. The final library size and quantity were verified using the FEMTO Pulse and the Qubit dsDNA HS Assay Kit. Sequencing primer and Sequel II DNA Polymerase were then annealed and bound to the SMRTbell library. The library was loaded at an on-plate concentration of 120 pM using diffusion loading. SMRT sequencing was conducted on a single 8 M SMRT Cell using the Sequel II System and the Sequel II Sequencing Kit, with 1800-minute movies produced by Frasergen Bioinformatics Co., Ltd. (Wuhan, China), and the resulting data was processed using SMRT LINK 5.0 software<sup>26</sup>.

**Genome assembly and analysis.** SMRTbell libraries were sequenced on a PacBio Revio system, and consensus HiFi reads were generated using the CCS software (<https://github.com/pacificbiosciences/unanimity>) with a minimum of three passes (‘-minPasses 3’). To further enhance and validate the assemblies, we generated 35.75 Gb of PacBio HiFi reads for this sample (Table 1), and the length distribution is shown in Fig. 2. These long (~15 kb), highly accurate (>99%) HiFi reads were assembled using Hifiasm (<https://github.com/chhylp123/hifiasm>)<sup>27</sup>. To correct errors in the initial assembly, Illumina-derived short reads were employed using Pilon (v1.23)<sup>28</sup> for error correction. The final *A. medius* genome assembly had a total length of approximately 739.07 Mb, consisting of 697 contigs with a Contig N50 of 10.004 Mb (Table 2). The completeness of the assembled genome was

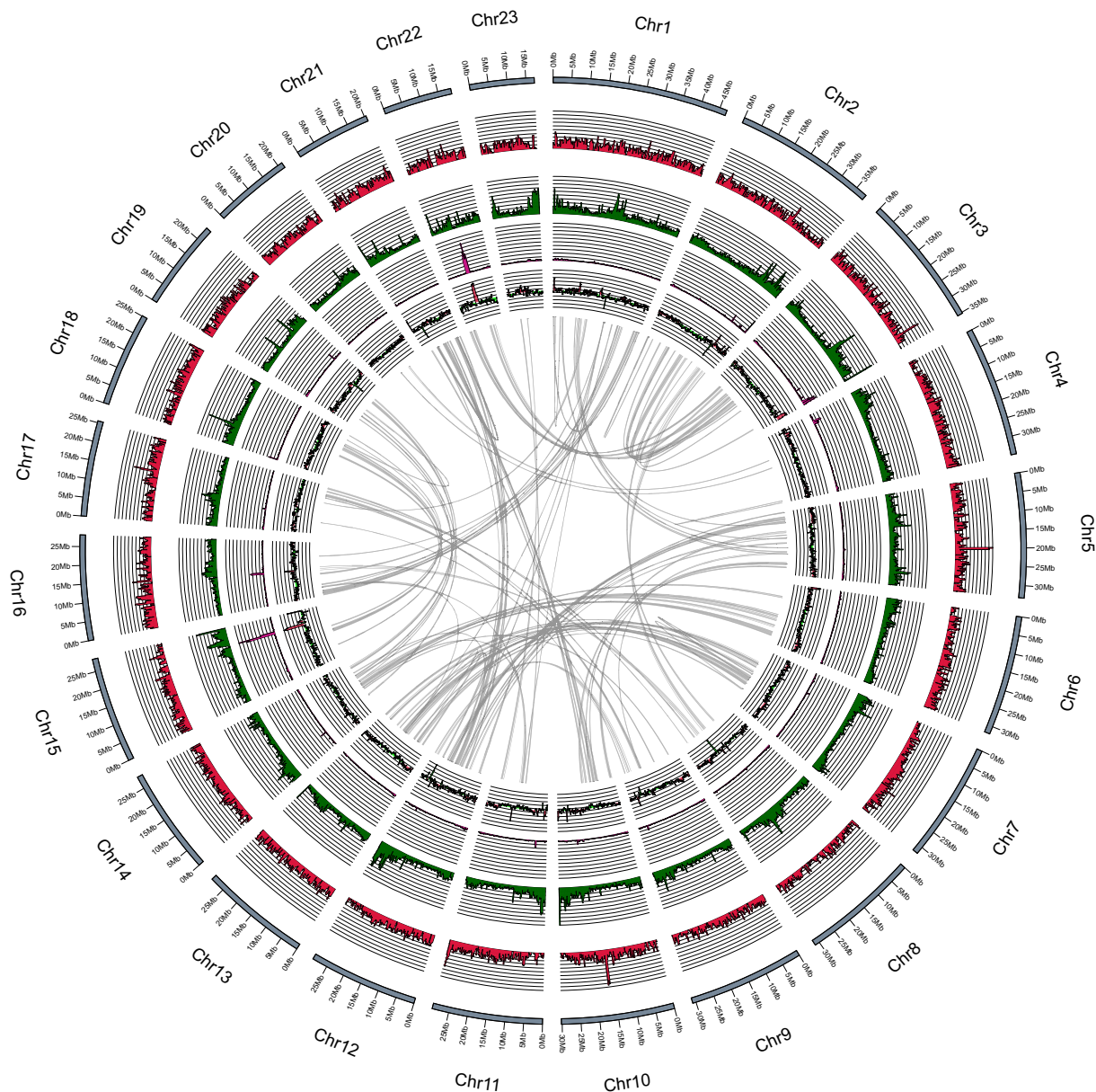
Repeat type	Repeat size (bp)
Low_complexity	2,247,372
LINE/L1-Tx1	163,046
LTR/DIRS	423,042
DNA/TcMar-ISRM11	232,981
DNA/TcMar-Tc1	2,091,901
DNA/hAT-Charlie	6,255,762
Unknown	165,243,169
DNA/hAT-Ac	1,456,899
LTR/ERV1	772,382
DNA/PIF-Harbinger	60,382
LTR/ERV-Foamy	144,344
DNA/hAT-Tol2	751,286
DNA/PiggyBac	479,504
DNA/hAT-hAT5	149,117
LINE/L1	1,736,919
LINE/Proto2	240,675
DNA/Academ	249,070
LTR/Pao	432,506
LINE	65,089
DNA/CMC-EnSpm	16,213
DNA/hAT-Tip100	4,163,103
LINE/L2	5,271,186
DNA/hAT-Blackjack	57,277
LTR/Gypsy	1,853,475
LINE/Rex-Babar	4,963,062
LINE/RTE-X	601,957
LINE/Jockey	86,961
LINE/I-Nimb	1,263,675
LTR	97,679
RC/Helitron	33,654
LINE/RTE-BovB	2,203,609
Simple_repeat	36,787,604
LTR/Ngaro	237,122
Total	240,832,023

**Table 4.** The summary of repeat sequence statistics of *A. medius* genome.

assessed using BUSCO (v3.1.0) with the OrthoDB database *Actinopterygii\_odb9*<sup>29,30</sup>, where 97.05% of the BUSCO genes were successfully identified (Fig. 3), indicating a highly complete genome assembly.

To anchor the contigs, 7.39 Gb of clean read pairs were generated from the Hi-C library and mapped to the polished *A. medius* genome using Bowtie 2 (v2.4.1) with default parameters<sup>31</sup>. Paired reads, where mates were mapped to different contigs, were utilized for Hi-C-assisted scaffolding. Invalid reads, such as self-ligation, non-ligation, and others like Start Near Rsite, PCR amplification artifacts, random breaks, Large Small Fragments, and Extreme Fragments, were filtered out using the HICUP software<sup>32</sup>. We clustered 390 contigs into 23 groups using agglomerative hierarchical clustering within 3D-DNA, as illustrated in Fig. 4. The 3D-DNA tool was also employed to order and orient these clustered contigs<sup>33,34</sup>. A total of 285 contigs, spanning 669,270,049 base pairs, were successfully ordered and oriented. This effort resulted in the first chromosomal-level high-quality assembly, with chromosome lengths ranging from 17.25 Mb to 46.97 Mb, accounting for 90.53% of the total genome sequence (Table 2). The statistics of *A. medius* chromosome lengths are summarized in Table 3.

**Repetitive sequence annotation and analysis.** To identify the repeat elements in *A. medius* genome, we applied two complementary methods: homology-based and de novo prediction. For the homology-based approach, we used RepeatMasker (v4.0.9)<sup>35</sup> with the Repbase TE library to identify known transposable elements (TEs). In parallel, we constructed a de novo repeat library for the *A. medius* genome using RepeatModeler (<http://www.repeatmasker.org/RepeatModeler/>)<sup>36</sup>, which automatically integrates two key de novo repeat-finding tools: RECON (v1.08) and RepeatScout (v1.0.5). These tools enabled comprehensive identification, refinement, and classification of consensus models for potential interspersed repeats<sup>37</sup>. Additionally, a de novo search for long terminal repeat (LTR) retrotransposons was conducted using LTR\_FINDER (v1.0.7), LTR harvest (v1.5.11), and LTR retriever (v2.7)<sup>38–40</sup>. We also identified tandem repeats using the Tandem Repeat Finder (TRF) package and simple sequence repeats (SSR) using MISA (v1.0)<sup>41</sup>. Finally, we merged the libraries generated from both approaches and used RepeatMasker<sup>35</sup> to identify the complete repeat content in the genome. As a result, we

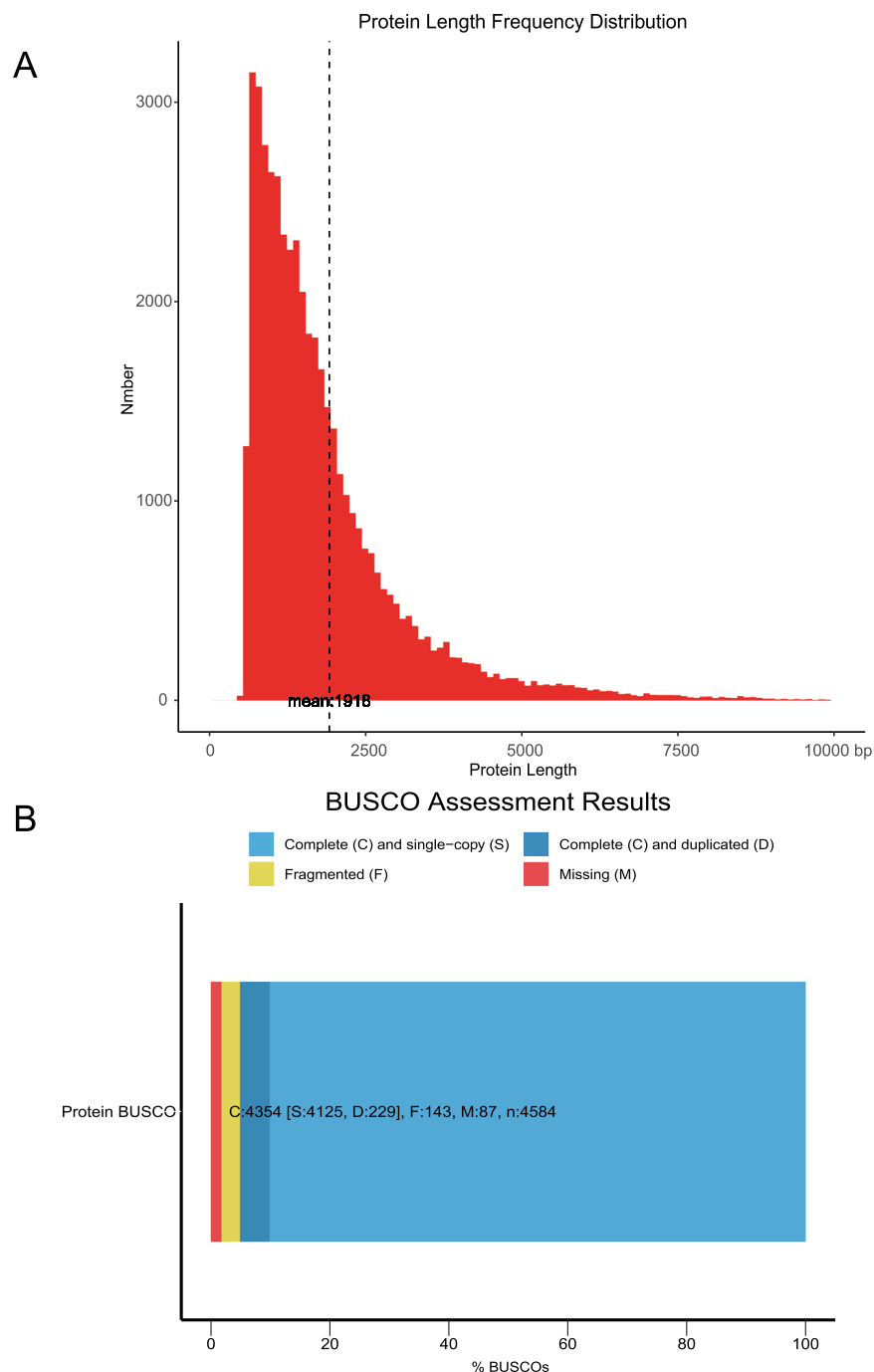


**Fig. 5** The genomic circus plot of *A. medius*. The rings, from the outermost to the innermost ring, the map illustrates the following: gene density (using a 200 kb sliding window), transposon density (200 kb sliding window), repetitive sequence density (marked as SimpleRepeat by RepeatMasker, with a 200 kb window), and GC content (200 kb sliding window, where red indicates regions with higher GC content than the genome average, and green indicates regions with lower GC content), the innermost ring represents collinear blocks.

rRNA type	Total length (bp)	Average length (bp)	Number
18s_rRNA	104,595	1,835	57
28s_rRNA	313,648	5808.296296	54
5_8S_rRNA	9,028	153.0169492	59
5S_rRNA	491,254	117.9198272	4,166
8s_rRNA	67,482	113.9898649	592

**Table 5.** The summary of non-coding RNA statistics of *A. medius* genome.

identified 240.83 Mb of repeat sequences within *A. medius* genome. A summary of these repeat sequence statistics is provided in Table 4. Furthermore, The Circos program<sup>42</sup> was used to generate a circular genome map, displaying various genomic features in concentric layers. From the outermost to the innermost ring, the map illustrates the following: gene density (using a 200 kb sliding window), transposon density (200 kb sliding window),



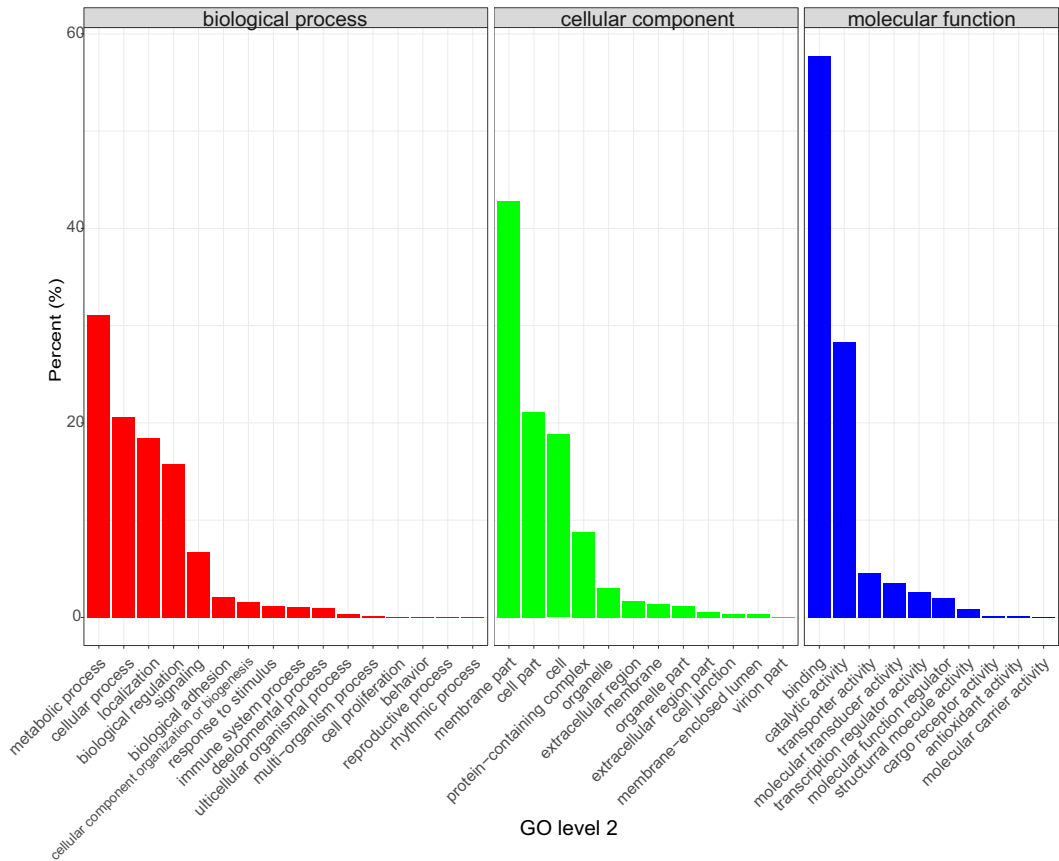
**Fig. 6** The protein-coding gene and protein BUSCO analysis for the sample of *A. medius*. **(A)** The protein length distribution. **(B)** The protein-coding gene BUSCO assessment for *A. medius* genome. Indicator: C, indicates complete BUSCOs; S, single-copy BUSCOs; D, complete and duplicated BUSCOs; F, fragmented BUSCOs; and M, missing BUSCOs; n, total number of BUSCO groups analyzed.

repetitive sequence density (marked as SimpleRepeat by RepeatMasker, with a 200 kb window), and GC content (200 kb sliding window, where red indicates regions with higher GC content than the genome average, and green indicates regions with lower GC content), the innermost ring represents collinear blocks (Fig. 5).

**Non-coding annotation and analysis.** We employed the tRNAscan-SE (v1.3.1)<sup>43</sup> algorithm with default parameters to identify tRNA genes. tRNAs are adaptor molecules composed of RNA that play a crucial role in translating the three-letter genetic code in messenger RNA (mRNA) into the twenty-letter code of amino acids in proteins. For rRNA gene identification, we used RNAmmer (v1.2)<sup>44</sup> with the parameters '-S euk -m lsu,ssu,tsu'. Additionally, snoRNAs, a class of small RNA molecules that guide the chemical modification of other RNAs (primarily ribosomal RNAs, transfer RNAs, and small nuclear RNAs), were identified using cmscan

Database	Annotated number	Annotated ratio
GO	22,336	47.93%
KEGG	26,480	56.82%
KOG	28,526	61.21%
Nr	45,236	97.07%
Pfam	34,911	74.91%
SwissProt	37,252	79.93%
TrEMBL	44,852	96.24%
Total	45,380	97.38%

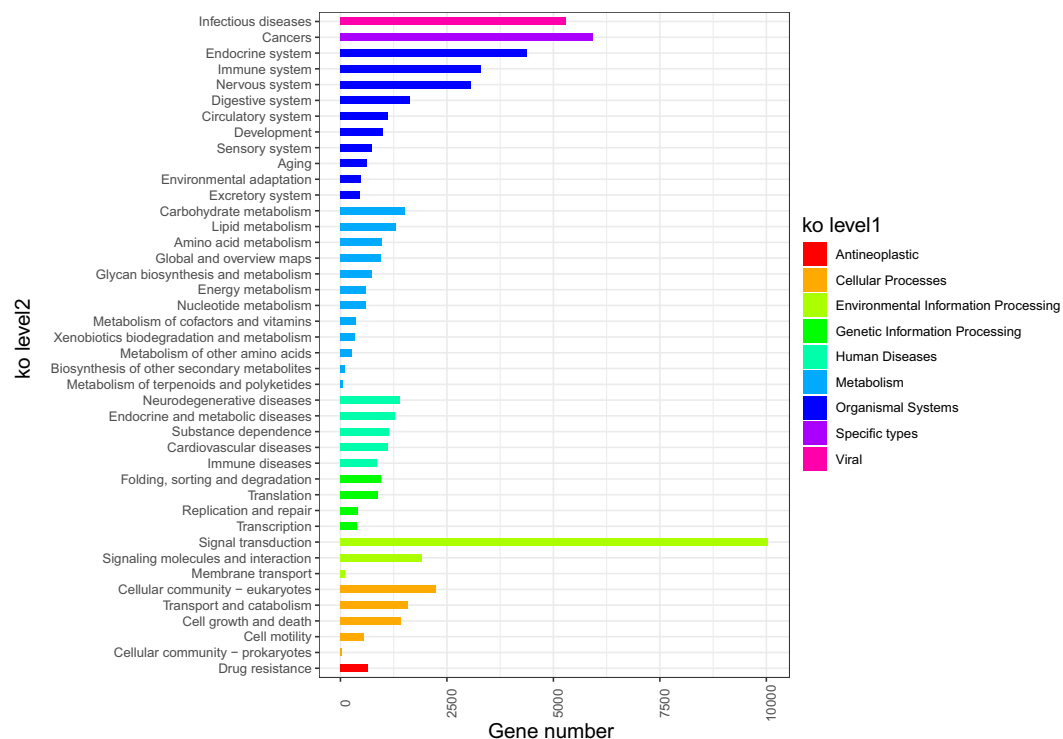
**Table 6.** The summary of protein-coding gene annotation statistics of *A. medius* genome.



**Fig. 7** Protein-coding gene annotations for Gene Ontology (GO) in *A. medius* genome.

(v1.1.2)<sup>45</sup> against the Rfam (v14.0)<sup>46</sup> database with default parameters. In the rRNA analysis, we identified a total of 4,928 rRNA sequences, with lengths ranging from 9,028 bp to 491,254 bp (Table 5). For tRNAs, 9,663 sequences were identified, with a total length of 728,112 bp and an average length of 75 bp, accounting for 0.0011 of the genomes. Additionally, we identified 347 snRNAs and 21 snoRNAs, with total lengths of 45,796 bp and 2,661 bp, respectively.

**Protein coding gene prediction and analysis.** For protein-coding genes prediction in *A. medius* genome using three complementary methods: ab initio gene prediction, homology-based prediction, and RNA-Seq-guided prediction. Before gene prediction, the assembled *A. medius* genome underwent both hard and soft masking with RepeatMasker<sup>35</sup> to eliminate repetitive elements. For ab initio gene prediction, we used Augustus (v3.3.3)<sup>47,48</sup>, which was trained with high-quality protein models generated from RNA-Seq data. For homology-based gene prediction, we employed MAKER (v2.31.10)<sup>49</sup>. In this process, protein and transcript sequences were aligned to the genome assembly, and coding genes were predicted using the default parameters in MAKER. RNA-Seq-guided gene prediction involved aligning clean RNA-Seq reads to the genome using HISAT2 (v2.0.0)<sup>50</sup>. The gene structures were then reconstructed using Trinity (v2.3.2)<sup>51</sup> with the parameter ‘-genome\_guided\_bam XXX\_aligned.bam’, followed by TransDecoder (v2.01) and MAKER<sup>49</sup> to further refine the gene models. Finally, the gene prediction results from the three methods were integrated using EVidenceModeler



**Fig. 8** Protein-coding gene annotations for KEGG in *A. medius* genome.

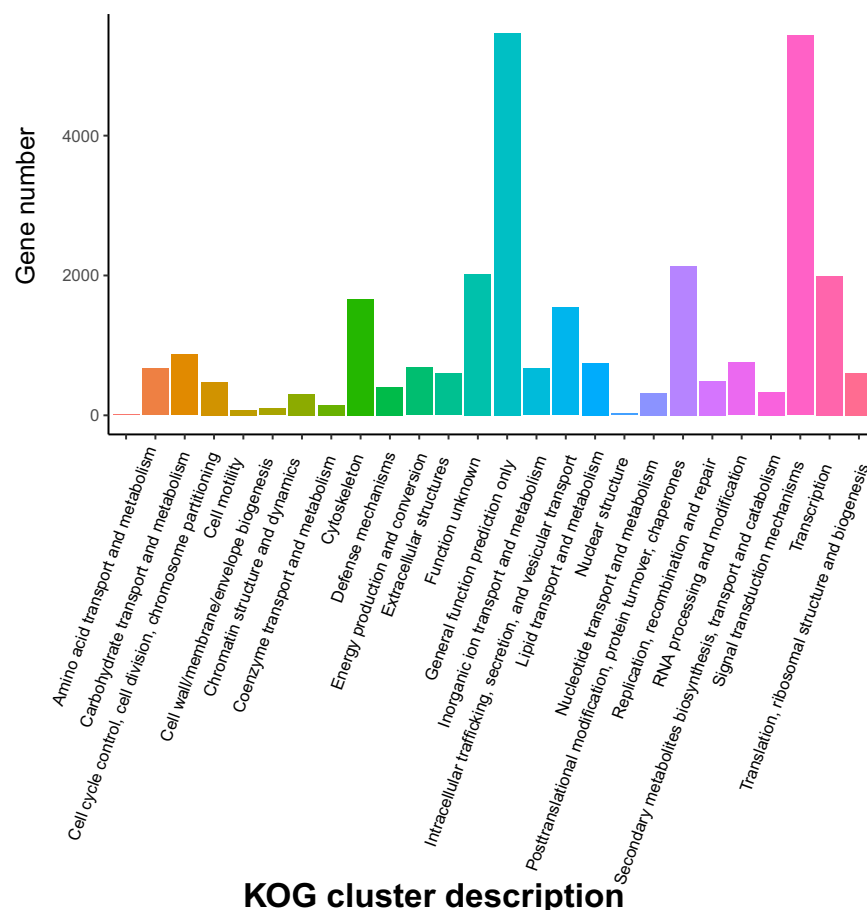
(EVM, v1.1.1)<sup>52</sup>, resulting in a consistent set of non-overlapping sequence assemblies that defined the gene structures. In total, 30,345 protein-coding genes were predicted, with an average gene length of 8,563.51 bp. Additionally, 46,603 proteins were identified, with a total length of 23,509,360 amino acids (aa) and an average length of 504.46 aa. Figure 6A illustrates the protein length distribution for the *A. medius* genome. The completeness of the predicted protein-coding genes was assessed using BUSCO, with the *Actinopterygii* odb9 database<sup>29,30</sup>, and the protein-coding genes datasets showed a 94.98% completeness score (Fig. 6B).

**Protein coding gene annotation and analysis.** Gene functions were predicted by aligning sequences to several major protein databases, including the National Center for Biotechnology Information (NCBI) Non-Redundant (NR), TrEMBL, InterPro<sup>53</sup>, and Swiss-Prot<sup>54</sup>, using BLASTP (ncbi-blast v2.6.0+)<sup>55</sup> with an E-value threshold of  $1E^{-5}$ . Additionally, functional annotation was performed using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database<sup>56</sup>. Protein domain identification was carried out using PfamScan<sup>57</sup> and InterProScan (v5.35–74.0)<sup>58</sup>, with reference to the InterPro protein database<sup>53</sup>. Motifs and domains within gene models were identified through the PFAM database<sup>59</sup>. Gene Ontology (GO) terms<sup>60</sup> were assigned to each gene using Blast2GO<sup>61</sup>. In total, approximately 45,380 (about 97.38%) of the predicted protein-coding genes in *A. medius* were successfully annotated with known gene functions, conserved domains, and GO terms. Table 6 provides a summary of the annotation statistics across the seven major databases.

In the GO annotation, the molecular function category showed the highest rate of annotations for binding (57.75%), followed by catalytic activity (28.28%). In the cellular component category, membrane parts had the most annotations (42.82%), followed by cell parts (21.12%). For the biological process category, metabolic processes (31.10%) and cellular processes (20.59%) were the top two with the highest annotation rates (Fig. 7). In the KEGG annotation, within the organismal systems category, the endocrine system had the highest number of annotated genes (4,368), followed by the immune system (3,291) and nervous system (3,075). Carbohydrate metabolism, lipid metabolism, and amino acid metabolism also showed a significant number of annotated genes, with 1,510, 1,300, and 980 genes, respectively. In the genetic information processing category, folding, sorting, and degradation were associated with 959 genes, while translation had 870 annotated genes. For environmental information processing, 10,047 genes were annotated under signal transduction, and 1,912 under signaling molecules and interaction. In the cellular processes category, cellular communities-eukaryotes, and transport and catabolism had 2,238 and 1,593 annotated genes, respectively. Figures 8 and 9 illustrate the protein-coding gene annotations from the KEGG and KOG databases.

### Data Records

All data from *A. medius* whole-genome project have been submitted to the NCBI SRA database under BioProject accession number PRJNA1176308 (<https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1176308/>). Specifically, the genome PacBio sequencing data was deposited at GeneBank (accession SRR31143456)<sup>62</sup>. The genome Hi-C sequencing data was deposited at GeneBank (accession SRR31715051)<sup>63</sup>. The RNA PacBio sequencing data



**Fig. 9** Protein-coding gene annotations for KOG in *A. medius* genome.

was deposited at GeneBank (accession SRR31715464)<sup>64</sup>. And The RNA BGI sequencing data was deposited at GeneBank (accession SRR31715465)<sup>65</sup>. This Whole Genome Shotgun project has been deposited at GenBank under the accession JBJXUX000000000<sup>66</sup>. Moreover, the assembly of *A. medius* has been deposited at GenBank under the accession GCA\_046127435.1 ([https://www.ncbi.nlm.nih.gov/datasets/genome/GCA\\_046127435.1/](https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_046127435.1/)). Concurrently, the final assembly and annotation files of this study are publicly available in FigShare (<https://doi.org/10.6084/m9.figshare.27368472>)<sup>67</sup>.

### Technical Validation

To ensure the accuracy of sequencing data, we performed a series of quality control checks on the samples before proceeding with library construction. First, we assessed the purity of the genomic DNA using a Nanodrop spectrophotometer (NanoDrop Technologies, DE, USA), measuring the OD260/280 and OD260/230 ratios and ensuring that the nucleic acid absorption peaks were normal. Next, the concentration of the genomic DNA was accurately measured using a Qubit fluorometer (Life Technologies, CA, USA), and the Qubit concentration was compared with the Nanodrop concentration to evaluate the sample's purity. Finally, the integrity of the genomic DNA was verified through electrophoresis. Only after meeting all quality standards were the samples used for library preparation. The completeness of the assembled genome was evaluated using BUSCO (v3.1) analysis with the OrthoDB database. The results showed that 97.05% of the BUSCO genes (*Actinopterygii* odb9) were successfully identified, indicating a high level of completeness in the assembled genome (Fig. 2).

### Code availability

No custom scripts or codes were used in the management and verification of the data sets in this study. All software and pipelines used for data processing were executed according to the manuals and protocols of the bioinformatics software cited above. The specific parameters were described if the default parameters were not applied for data analysis.

Received: 25 November 2024; Accepted: 3 January 2025;  
Published online: 14 January 2025

## References

- Mecklenburg, C. W., Lynghammar, A., Johannesen, E., Byrkjedal, I. & Wienerroither, R. M. 2018. Marine Fishes of the Arctic Region. *Conservation of Arctic Flora and Fauna*. 372–373 (2018).
- Liu, K., Lin, H., Wang, J. & Sun, J. Complete mitochondrial genome sequence and phylogenetic analysis of *Anisarchus medius* (Reinhardt, 1837). *Mitochondrial DNA B*. **4**(2), 3973–3974 (2019).
- Whitehead, P., Bauchot, M., Hureau, J., Nielsen, J. & Tortonese, E. *Fishes of the North-eastern Atlantic and the Mediterranean*. Vol. 3 1126–1129 (Springer, 1986).
- McMeans, B. C., Rooney, N., Arts, M. T. & Fisk, A. T. Food web structure of a coastal Arctic marine ecosystem and implications for stability. *Mar. Ecol. Prog. Ser.* **482**, 17–28 (2013).
- Rantanen, M. *et al.* The Arctic has warmed nearly four times faster than the globe since 1979. *Commun Earth Environ.* **3**, 168 (2022).
- Post, E. *et al.* Ecological dynamics across the Arctic associated with recent climate change. *Science*. **325**(5946), 1355–1358 (2009).
- Johannessen, O. M. & Miles, M. W. Critical vulnerabilities of marine and sea ice-based ecosystems in the high Arctic. *Reg. Environ. Change*. **11**(Suppl 1), 239–248 (2011).
- Forster, C. E., Norcross, B. L. & Spies, I. Documenting growth parameters and age in Arctic fish species in the Chukchi and Beaufort seas. *Deep Sea Research Part II: Topical Studies in Oceanography* **177** (2020).
- Looby, A. *et al.* Global inventory of species categorized by known underwater sonifery. *Sci Data* **10**, 892 (2023).
- Voskoboinikova, O. Skeleton development in ontogenesis of the butterfish *Pholis gunnellus*, *Anisarchus Anisarchus medius*, and slender eel blenny *Lumpenus fabricii* (Zoarcoidei, Perciformes). *J Ichthyol* **45**, 503–512 (2005).
- Chiperzak, D., Hopky, G., Lawrence, M., Schmid, D. & Reist, J. *Larval and post-larval fish data from the Canadian Beaufort Sea shelf, July to September, 1987*. (Fisheries and Oceans Canada Winnipeg MB, 2003).
- Turanov, S., Kartavtsev, Y. P. & Zemnukhov, V. Molecular phylogenetic study of several eelpout fishes (Perciformes, Zoarcoidei) from Far Eastern seas on the basis of the nucleotide sequences of the mitochondrial cytochrome oxidase 1 gene (CO1). *Russ. J. Genet.* **48**, 208–223 (2012).
- Dahl, T. M. *et al.* Fatty acid composition of the blubber in white whales (*Delphinapterus leucas*). *Polar Biol.* **23**, 401–409 (2000).
- Murzina, S. A. *et al.* Lipid status of the two high latitude fish species, *Leptoclinus maculatus* and *Lumpenus fabricii*. *Int J Mol Sci.* **14**, 7048–7060 (2013).
- Prosser, C. L., Mackay, W. & Kato, K. Osmotic and ionic concentrations in some Alaskan fish and goldfish from different temperatures. *Physiol. Zool.* **43**, 81–89 (1970).
- Lin, L. *et al.* Composition and distribution of fish species collected during the fourth Chinese National Arctic Research Expedition in 2010. *Arctic Research* **23**, 116 (2012).
- Norcross, B. L., Holladay, B. A., Busby, M. S. & Mier, K. L. Demersal and larval fish assemblages in the Chukchi Sea. *Deep Sea Research Part II: Topical Studies in Oceanography* **57**, 57–70 (2010).
- Frainer, A. *et al.* Climate-driven changes in functional biogeography of Arctic marine fish communities. *Proceedings of the National Academy of Sciences* **114**, 12202–12207 (2017).
- Fu, S. *et al.* Trophic levels and trophic niches of fish from the Bering Sea and Chukchi Sea. *Biodiversity Science* **31**, 22521 (2023).
- Ghezelayagh, A. *et al.* Prolonged morphological expansion of spiny-rayed fishes following the end-Cretaceous. *Nat Ecol Evol* **6**, 1211–1220 (2022).
- Mu, Y. *et al.* Whole genome sequencing of a snailfish from the Yap Trench (~7,000 m) clarifies the molecular mechanisms underlying adaptation to the deep sea. *PLoS Genet.* **17**, e1009530 (2021).
- Liu, R. *et al.* De novo genome assembly of limpet *Bathymacra lactea* (Gastropoda: Pectinodontidae), the first reference genome of a deep-sea gastropod endemic to cold seeps. *Genome Biol. Evol.* **12**, evaa100 (2021).
- Liu, R., Liu, J. & Zhang, H. Positive selection analysis reveals the deep-sea adaptation of a hadal sea cucumber (*Paelopatides* sp.) to the Mariana Trench. *J. Oceanol. Limnol.* **39**, 266–281 (2021).
- Feng, C. *et al.* The genome of a new anemone species (Actiniaria: Hormathiidae) provides insights into deep-sea adaptation. *Deep-Sea Res. I: Oceanogr. Res. Pap.* **170**, 103492 (2021).
- Ao, J. *et al.* Genome sequencing of the perciform fish *Larimichthys crocea* provides insights into molecular and genetic mechanisms of stress adaptation. *PLoS Genet.* **11**, e1005118 (2015).
- Flusberg, B. *et al.* Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Methods* **7**, 461–465 (2010).
- Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
- Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*. **9**(11), e112963 (2014).
- Manni, M., Berkeley, M. R., Seppely, M. & Zdobnov, E. M. BUSCO: assessing genomic data quality and beyond. *Curr. Protoc.* **1**(12), e323 (2021).
- Kuznetsov, D. *et al.* OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Res.* **51**(D1), D445–51 (2023).
- Li, H. & Durbin, R. Fast and accurate long-read alignment with burrows–wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
- Wingett, S. *et al.* HiCUP: pipeline for mapping and processing Hi-C data. *F1000Res.* **4**, 1310 (2015).
- Servant, N. *et al.* HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 1–11 (2015).
- Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst.* **3**, 95–98 (2016).
- Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinform.* **25**(1), 4–10 (2009).
- Flynn, J. M. *et al.* RepeatModeler 2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA* **117**, 9451–9457 (2020).
- Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
- Xu, Z. & Wang, H. LTR\_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
- Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinform.* **9**, 18 (2008).
- Ou, S. & Jiang, N. LTR\_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).
- Beier, S., Thiel, T., Münch, T., Scholz, U. & Mascher, M. MISA-web: a web server for microsatellite prediction. *Bioinformatics*. **33**(16), 2583–2585 (2017).
- Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**(9), 1639–1645 (2009).
- Lowe, T. M. & Eddy, S. R. tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
- Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35**, 3100–8 (2007).
- Nawrocki, E. P. & Eddy, S. C. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
- Kalvari, I. *et al.* Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res.* **49**(D1), D192–200 (2021).

47. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**(5), 637–644 (2008).
48. Stanke, M., Schöffmann, O., Morgenstern, B. & Waack, S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*. **7**, 1–11 (2006).
49. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatic*. **12**, 491 (2011).
50. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*. **12**(4), 357–60 (2015).
51. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protocols*. **8**(8), 1494–1512 (2013).
52. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EvidenceModeler and the program to assemble spliced alignments. *Genome Biol.* **9**, R7 (2008).
53. Mitchell, A. *et al.* The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.* **43**(D1), D213–21 (2015).
54. Boeckmann, B. *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**(1), 365–370 (2003).
55. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**(17), 3389–3402 (1997).
56. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **40**(D1), D109–114 (2012).
57. Mistry, J., Bateman, A. & Finn, R. D. Predicting active site residue annotations in the Pfam database. *BMC Bioinformatic*. **8**, 1–4 (2007).
58. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics*. **30**(9), 1236–1240 (2014).
59. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res.* **42**(D1), D222–230 (2014).
60. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**(1), 25–29 (2000).
61. Conesa, A. & Götz, S. Blast2GO: a comprehensive suite for functional analysis in plant genomics. *Int. J. Plant Genomics*. **1**, 619832 (2008).
62. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR31143456> (2024).
63. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR31715051> (2024).
64. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR31715464> (2024).
65. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR31715465> (2024).
66. Liu, R. Y. *Anisarchus medius*, whole genome shotgun sequencing project. *GenBank* <https://identifiers.org/ncbi/insdc:JBXUX000000000> (2024).
67. Liu, R. Y. Genome annotations of *Anisarchus medius*. *Figshare dataset* <https://doi.org/10.6084/m9.figshare.27368472> (2024).

## Acknowledgements

This research was supported by the National Natural Science Foundation of China (grants 42276247 and 32202995), National Key Research and Development Program of China (2022YFD2401001), and Natural Science Foundation of Fujian Province (2022J01135). We would like to express our sincere gratitude to the captain, crew, and scientific expedition members of the icebreaker R/V Xuelong (Snow Dragon), especially Pengfei Cui from the Third Institute of Oceanography, Ministry of Natural Resources, for their support and cooperation during the 2014 Chinese Arctic Research Expedition. We also appreciate Ke Xing and Jianan Chen for their assistance in sample preservation, genome survey and manuscript revision. Additionally, we are indebted to the Chinese Scientific Committee on Oceanic Research for providing an invaluable platform for academic discussion and exchange.

## Author contributions

Xinhua Chen, Longshan Lin, and Ruoyu Liu conceived of the project; Scientific expedition members of the icebreaker R/V Xuelong (Snow Dragon), especially Pengfei Cui collected the samples; Ruoyu Liu and Yanan Mu preserved the samples. Longshan Lin and Ran Zhang collected photos of this species. Ruoyu Liu and Ziyu Meng extracted the genomic DNA; Ruoyu Liu and Ping Zheng performed the data analysis and wrote the manuscript; Ruoyu Liu revised the manuscript. All authors read and approved the final version of the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to L.L., P.Z. or X.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-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, 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 you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. 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-nc-nd/4.0/>.

© The Author(s) 2025