Scientific Data

https://doi.org/10.1038/s41597-026-07126-1

Article in Press

An annotated genome of freshwater amphipod (*Gammarus nekkensis*)

Received: 10 October 2025

Accepted: 25 March 2026

Cite this article as: Lu, D., Liu, H., Tong, Y. *et al.* An annotated genome of freshwater amphipod (*Gammarus nekkensis*). *Sci Data* (2026). https://doi.org/10.1038/s41597-026-07126-1

Decai Lu, Hongguang Liu, Yan Tong, Zeyu Liu, Chao-Dong Zhu & Zhonge Hou

An annotated genome of freshwater amphipod (*Gammarus nekkensis*)

Decai Lu[1,2], Hongguang Liu[1], Yan Tong[1,2], Zeyu Liu[1,2], Chao-Dong Zhu[1] & Zhonge Hou[3*]

[1]State Key Laboratory of Animal Biodiversity Conservation and Integrated Pest Management, Institute of Zoology, Chinese Academy of Sciences, 100101 Beijing, China

[2]University of Chinese Academy of Sciences, 100049 Beijing, China

[3]College of Life Sciences, Capital Normal University, 100048 Beijing, China

[*]email: houze@cnu.edu.cn

Abstract

Freshwater *Gammarus* species represent important keystone organisms in aquatic ecosystems, and are sensitive to environmental changes. Here, we present the first pseudo-chromosome-level genome of *Gammarus nekkensis*, endemic to north China. We integrated PacBio HiFi long-read sequencing, Illumina short-read sequencing, and Hi-C scaffolding to generate a high-quality, pseudo-chromosome-scale genome assembly. The assembled genome is approximately 6.24 Gb in size, with a scaffold N50 of 233.63 Mb, and 96.76% of the sequences were successfully anchored to 26 pseudo-chromosomes. A total of 39,474 protein-coding genes were predicted, and approximately 70% of these genes obtained functional annotations. Repetitive elements constituted about 63.93% of the genome, with long interspersed nuclear elements (LINE) being the most abundant (23.98%). BUSCO analysis indicated that both the assembly and annotation are highly complete compared with published amphipod genomes. This high-quality genome enables studies of sex determination, adaptive evolution, and genomic diversity in *G. nekkensis*, with applications for its conservation and breeding.

Background & Summary

Environmental changes including pollution and climate warming are dramatically altering water systems and affecting species distribution[1,2]. Freshwater amphipods are widely distributed in northern hemisphere. They stand for important keystone species in aquatic ecosystems, due to their high abundance and biomass. Their population health is often used as a bioindicator for monitoring water quality and environmental pollution[3]. Amphipods are also significant components in nutrient recycling and provide high-quality food for fish and water birds[4]. In recent years, amphipods have gained prominence as model organisms for functional genomics and genome editing, especially in investigation of cell fate specification and developmental plasticity[5,6].

Species of *Gammarus* (Malacostraca; Amphipoda; Gammaridae) species are broadly distributed in alpine streams and boreal lakes, which are considered typical cold-water species[7,8]. Previous studies revealed that past climate cooling promoted diversification of freshwater *Gammarus* in Eurasia, however, future climate warming would influence its distribution[9]. The evolutionary trajectory and thermal sensitivity makes freshwater *Gammarus* species ideal models for studying genomic signature of cold adaption under climatic pressure. However, the absence of high-quality genomic resources has restricted investigations between genetic response and environment changes.

Here, we present the first pseudo-chromosome-level genome assembly of *G. nekkensis*, a cold-water amphipod species distributed in the surrounding areas of north China. It inhabits mountain streams at elevations about 2000 meters. This species feeds on decaying leaves and carrion, and exhibits a strong preference for hiding in light-avoided areas beneath rocky substrates. Based on k-mer analysis of Illumina short-read data, the genome

size of *G. nekkensis* is greater than 4.57 Gb (Fig. 1).

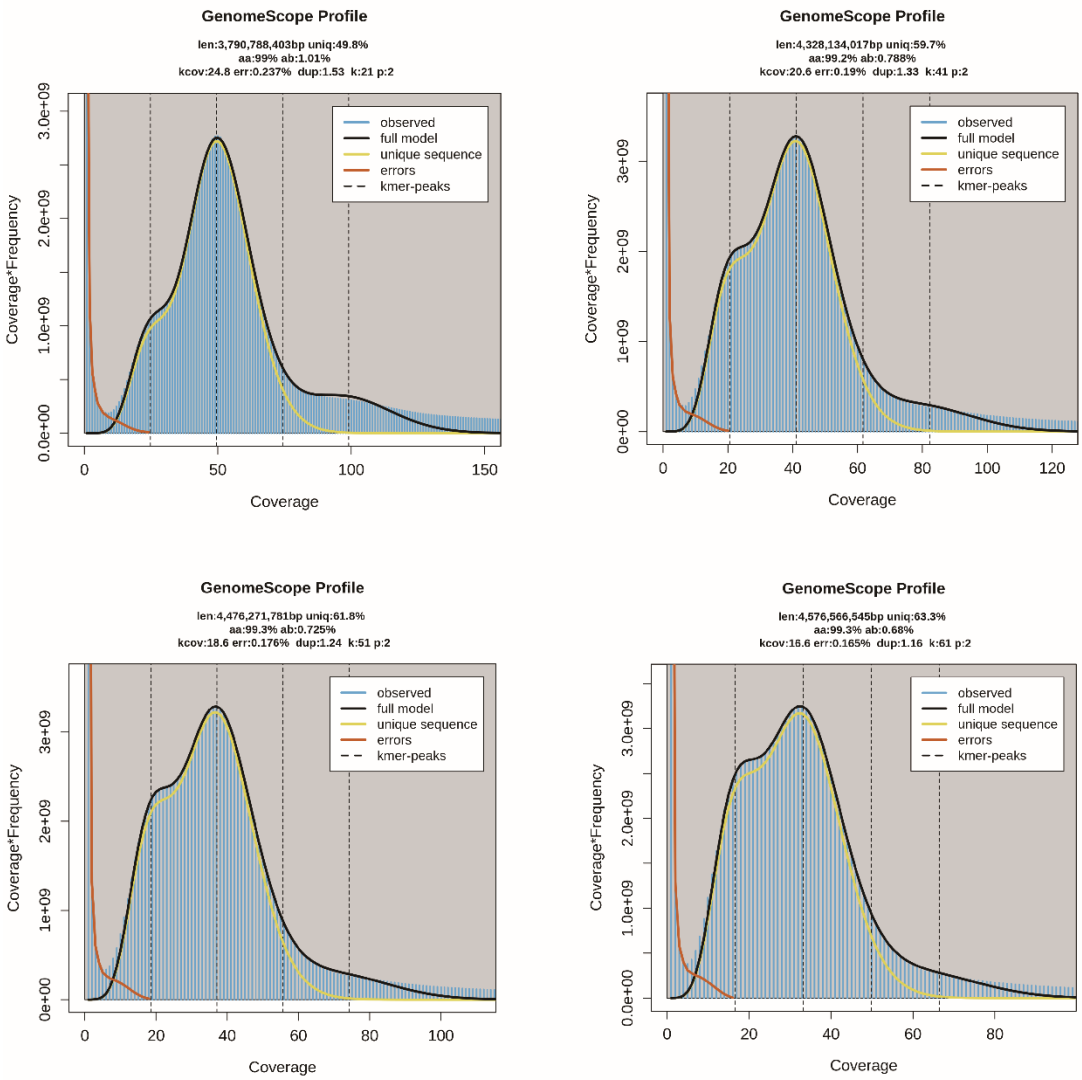

Fig. 1 Genome size estimation of *Gammarus nekkensis* using k-mer analysis with different k values. (a) k=21. (b) k=41. (c) k=51. (d) k=61.

Using PacBio CCS (134.41 Gb) and Hi-C chromatin conformation capture data (507.36 Gb), we assembled a 6.24 Gb genome, of which 96.7% (6.04 Gb) was anchored onto 26 chromosomal scaffolds (Fig. 2). This size exceeds initial estimates, which may be due to the abundance of repetitive sequences in the genome that affected the genome size assessment[10].

Among the publicly available Amphipoda genome assemblies, the genome size in this study is the second largest to that of *Hirondellea gigas* (13.92 Gb). Moreover, this genome represents the third pseudo-chromosome-level assembly published for amphipods (Table 1). By assessing the completeness and accuracy of these genome assemblies using the Benchmarking Universal Single-Copy Orthologs (BUSCO) v5.4.2[11] with the arthropod_odb10 dataset, we found this genome shows a relatively high score as 88.4% (Table 1).

Table 1 The information on genome assemblies in the public databases for Amphipoda.

Species	Genome size (Gb)	Complete BUSCOs (%)	Assembly level	Assembly Accession
Gammarus nekkensis	6.24	88.4%	Pseudo-chromosome	This study
Hirondellea gigas	13.92	89.93%	Chromosome	CNA0142381[12]
Morinoia aosen	1.99	94.50%	Chromosome	GCA_030386875.1[13]
Parhyale hawaiensis	2.75	90.30%	Scaffold	GCA_001587735.2[14]
Hyalella azteca	0.54	95.60%	Scaffold	GCA_000764305.4[15]
Trinorchestia longiramus	0.87	86.40%	Scaffold	GCA_006783055.1[16]
Gammarus roeselii	3.24	35.20%	Scaffold	GCA_016164225.1[17]
Potiberaba porakuara	0.79	75.90%	Scaffold	GCA_047292215.1[18]
Phronima sedentaria	1.07	7.70%	Scaffold	GCA_037179465.1[19]
Talorchestia martensii	0.67	84.50%	Scaffold	GCA_054095995.1[20]
Caprella mutica	0.90	89.90%	Scaffold	GCA_947561585.1[21]
Orchestia grillus	0.81	67.10%	Scaffold	GCA_014899125.1[22]

Transposable elements (TEs) sequences account for 63.93% of the genome assembly. Apart from unclassified sequences, long interspersed nuclear elements (LINE) were the largest category of transposable elements, representing 23.98% of the genome, followed by long terminal repeats (LTR), representing 5.72% of the genome (Table 2).

Table 2 Characterization of transposable elements in the *Gammarus nekkensis* genome.

Transposable elements	Percentage in Genome	Length (M)
SINE	0.11%	6.864
Penelope	0.05%	3.12
LINE	23.98%	1496.352
LTR	5.72%	356.928
DNA	2.14%	133.536
Rolling-circles	0.01%	0.624
Unclassified	27.49%	1715.376
Small RNA	0.86%	53.664
Satellites	0.04%	2.496
Simple repeats	3.59%	224.016
Low complexity	0.06%	3.744

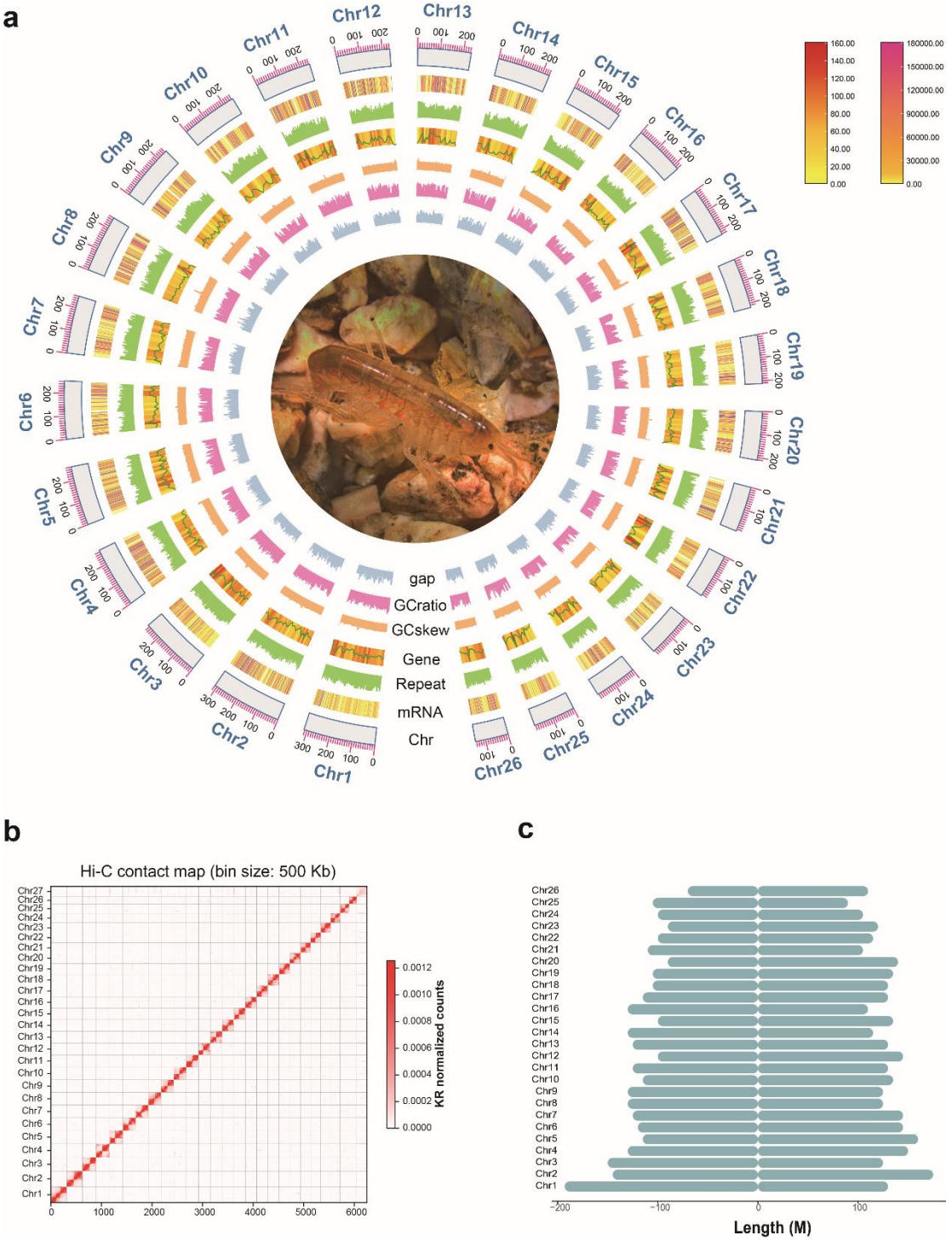

Fig. 2 Characterization of *Gammarus nekkensis* genome. (a) Circos plot of the genomic characteristics. From the inner to the outer ring, the tracks represent: genome assembly gaps, GC ratio, GC skew, gene density, repetitive sequences, transcriptional activity, and pseudo-chromosome numbers. (b) Hi-C contact map of the pseudo-chromosome-level assembly. (c) Graphical representation of pseudo-chromosome sizes and centromere locations.

Methods

Sample collection and sequencing

Individuals of *G. nekkensis* for genome assembly were collected from Hebei, China (40.62139°N, 117.43639°E) in June 2023. All samples were starved in the laboratory for one week to reduce contamination, and the feeding temperature was maintained at 16 ℃. Short-insert (350 bp) paired-end (PE) libraries were constructed using a Truseq DNA PCR-free kit with the Illumina NovaSeq 6000 platform. The largest individual was used to extract the total DNA for Whole Genome Sequencing (WGS) and PacBio High Fidelity Reads (HiFi) sequencing. The template preparation kit (PacBio) was used to prepare the PacBio Sequel II/IIe library. The DNA library was sequenced via PacBio Sequel II sequencers with the continuous long read model at Novogene (Beijing, China).

For the preparation of Hi-C libraries, cells were first cross-linked using 1% formaldehyde for ten minutes at room temperature, followed by quenching with glycine at a final concentration of 0.125 M for five minutes. After lysis of the cross-linked cells, nuclei were isolated and digested with 80 units of DpnII (NEB, R0543L) at 37 °C for four hours. The digested DNA was then biotin-labeled using biotin-14-dCTP (Invitrogen) and ligated with 4000 units of T4 DNA ligase (NEB) at 20 °C[23]. Cross-links were subsequently reversed, and the ligated DNA was purified using the QIAamp DNA Mini Kit (Qiagen) according to the manufacturer's protocol. The purified DNA was sheared into fragments ranging from 300 to 500 bp, subjected to blunt-end repair, A-tailing, and adapter ligation. Biotin-streptavidin pull-down was performed to enrich the target fragments, followed by PCR amplification. The final Hi-C libraries were quantified and sequenced on the Illumina NovaSeq 6000 platform[24].

Genome assembly

The distribution of k-mers from Jellyfish v2.2.0[25] (run with multiple k-mer sizes) was analyzed using GenomeScope v2.0[26] to estimate genome characteristics. Draft genome is assembled from high-quality HiFi long read data using Hifiasm v0.16.1[27]. Duplicate items were removed based on read depth with PurgeDups v1.2.5[28]. The draft assemblies were subsequently gap-filled using GapCloser v1.12[29] and polished with NextPolish v1.4.1[30]. Hi-C library sequencing data were mapped to the contig-level genome using BWA v0.7.18[31]. The Haphic v1.0.6[32] pipeline was executed to construct the chromosomal scaffolds and correct misassemblies. The final assembly files were generated using YaHS v1.2a.1[33]and refined with Juicebox Assembly Tools v1.9.9[34] (Fig. 3).

Genome annotation

The transposable elements sequences were annotated by a combination of homology-based and de novo approaches. Initially, a de novo repeat library based on the assembly sequences was generated by RepeatModeler v2.0.5[35] with default parameters. The library was used as the database for the identification of the TE sequences with homology searching by RepeatMasker v4.1.5[36] (Fig. 3).

Protein-coding gene ab-initio prediction was performed using Braker v3.0.8[37]. Homology-based gene prediction was conducted with GeMoMa v1.9[38], utilizing protein sequences from six reference species: *Hyalella azteca, Morinoia aosen, Eriocheir sinensis, Penaeus monodon, Drosophila melanogaster, Pollicipes pollicipes*. Finally, we integrated these two results into the final annotation through EVM[39] (Fig. 3).

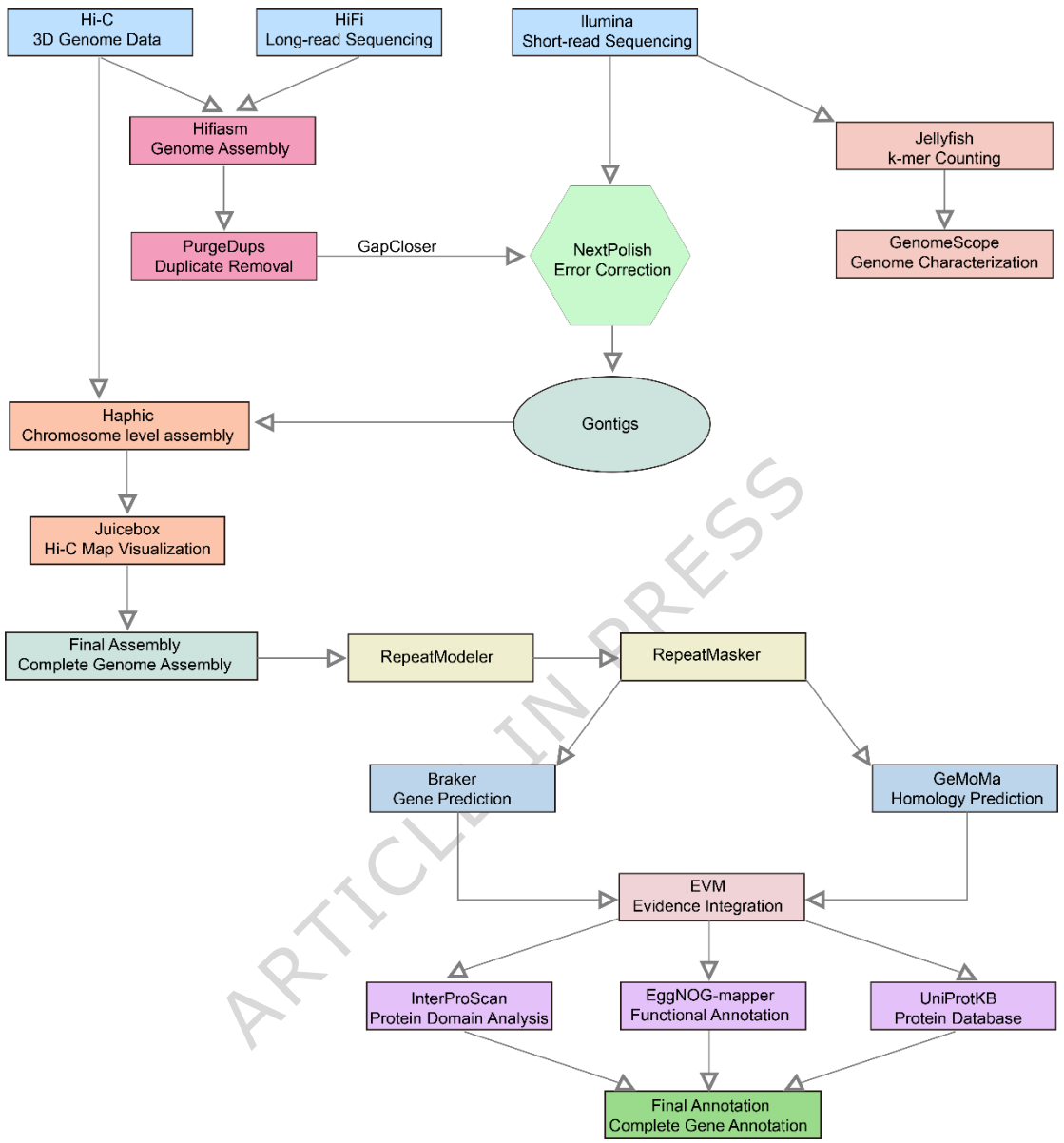

Fig. 3 The analytical flowchart of genome assembly and annotation in this study.

Functional assignment to protein sequences was conducted via alignment to the UniProtKB[40] database, utilizing the Diamond v0.9.24[41]. An integrated approach employing both eggNOG-mapper v2.1.12[42] and InterProScan v5.75[43] was utilized for the comprehensive functional annotation of Gene Ontology (GO) terms, Kyoto Encyclopedia of Genes and Genomes (KEGG) orthologs, and protein domains. The InterProScan analysis integrated data from five databases: Pfam, PRINTS, PANTHER, ProSiteProfiles and SMART. Chromosomal features, including repeat elements, gene density, transcript abundance distribution and GC content, were visualized using TBtools v2.332[44] (Fig. 3).

Data Records

The raw sequencing data supporting this study have been deposited in the European Nucleotide Archive (ENA) with the following accession numbers: PacBio data, ERR16720729[45], Illumina data, ERR16722743[46], and Hi-C data, ERR16729500[47]. The assembled genome sequence is available from the ENA under accession number ERZ29160471 (GCA_980912865)[48] and from the China National GeneBank under accession number CNA0509588[49]. The functional annotation information of the gene was uploaded to Figshare database under the following DOI: https://doi.org/10.6084/m9.figshare.31698019[50].

Genome Assembly: The final, pseudo-chromosome-level genome assembly is available in the file ERZ29091126. This assembly comprises 27 scaffolds in total. Among these, scaffolds 1 to 26 correspond to individual pseudo-chromosome, achieving pseudo-chromosome-scale continuity. The remaining sequences, primarily shorter contigs that could not be unambiguously placed, have been aggregated into a separate file for convenience, which is represented here as scaffold 27.

Gene annotation: The annotation information of protein-coding genes, along with the genome assembly data, was uploaded to the ENA in embl format. The functional annotation information of the gene was uploaded to Figshare[50].

Sequencing Data: The sequencing data generated for this project are derived from three complementary high-throughput sequencing platforms, each providing distinct and essential information for comprehensive genomic analysis. The files are systematically categorized into the following three primary types: (1) Illumina Second Generation Sequencing Data: This component consists of short-read data generated by Illumina platforms. (2) PacBio Third Generation Sequencing Data: This component comprises long-read data generated by Pacific Biosciences (PacBio) platforms. (3) Hi-C (Chromatin Conformation Capture) Sequencing Data: This component includes data obtained from Hi-C technology.

Data Availability Statement

The raw sequencing data supporting this study have been deposited in the European Nucleotide Archive (ENA) with the following accession numbers: PacBio data, ERR16720729, Illumina data, ERR16722743, and Hi-C data, ERR16729500. The assembled genome sequence is available from the ENA under accession number ERZ29160471 (GCA_980912865) and from the China National GeneBank under accession number CNA0509588. The functional annotation information of the gene was uploaded to Figshare database under the following DOI: https://doi.org/10.6084/m9.figshare.31698019.

Technical Validation

A dual-mode analytical approach was employed to assessing the completeness of the genome assembly by BUSCO. First, assembly continuity of genome was assessed through sequence mapping and Metaeuk-based gene prediction. Second, gene set completeness of protein was evaluated via alignment of predicted proteins to orthologous sequences. The results indicated that 74.6% of the conserved single-copy orthologs were present as complete, single-copy sequences. An additional 13.8% were classified as duplicated, which may indicate small-scale duplications or assembly artifacts. Among the remaining orthologs, 5.0% (51 genes) were fragmented, and 6.6% (66 genes) were entirely missing from the assembly (Table 3). By merging annotation results from ab-initio prediction and homology-based gene prediction, a total of 39,474 potential genes were annotated, among which 27,731 genes have functional annotation results (including 5,202 genes annotated as uncharacterized protein). The average amino acid length of the 27,731 functionally annotated genes is 501.84 bp (Supplementary

File 1). The completeness of protein annotation is consistent with genome conservation, approximately 87.6% (Table 3).

Table 3 The BUSCO assessment of *Gammarus nekkensis*.

Type	BUSCO of Genome	BUSCO of Annotation
Complete BUSCOs (C)	88.4% (896)	87.6% (888)
Complete and single-copy BUSCOs (S)	74.6% (756)	68.3% (692)
Complete and duplicated BUSCOs (D)	13.8% (140)	19.3% (196)
Fragmented BUSCOs (F)	5.0% (51)	4.2% (43)
Missing BUSCOs (M)	6.6% (66)	8.2% (82)
Total BUSCO groups searched	1013	1013

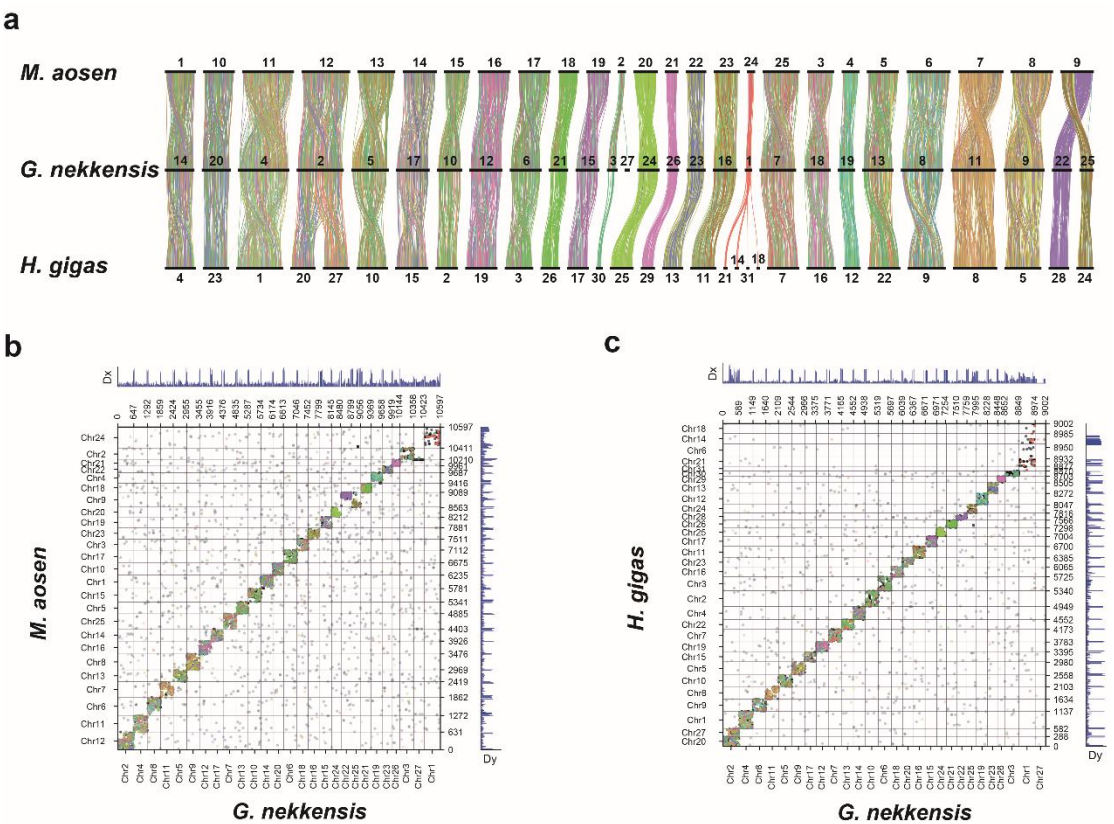

Fig. 4 Chromosomal collinearity between *Gammarus nekkensis* and reference species (*Morinoia aosen, Hirondellea gigas*). (a) The chromosomal synteny relationships among three species were visualized using the odp software suite. (b-c) Genome synteny circle plot of *Gammarus nekkensis* with two reference species.

To validate the quality of the pseudo-chromosome-level genome assembly, we performed whole-genome synteny visualization among three amphipod species using odp v0.3.3[51] software suite (Fig. 4). The results demonstrate that the newly assembled genome exhibits a high degree of stability and reliability.

Code availability

All analyses were conducted using the software and pipelines as specified in the Methods section, without the generation of any custom code.

Funding

This study was supported by the National Natural Science Foundation of China (grant number 32470474, 32500387), the International Partnership Program of Chinese Academy of Sciences (grant number 073GJHZ2024043MI), the Institute of Zoology, Chinese Academy of Sciences (2023IOZ0104, 2024IOZ0108); Beijing Natural Science Foundation (grant number 5244045).

Author information

Authors and Affiliations

[1]State Key Laboratory of Animal Biodiversity Conservation and Integrated Pest Management, Institute of Zoology, Chinese Academy of Sciences, 100101 Beijing, China
Decai Lu, Hongguang Liu, Yan Tong, Zeyu Liu, Chao-Dong Zhu

[2]University of Chinese Academy of Sciences, Beijing 100049, China
Decai Lu, Yan Tong, Zeyu Liu

[3]College of Life Sciences, Capital Normal University, Beijing 100048, China
Zhonge Hou

Contributions

D. L. and Z. H. conceived the study; D. L, Z. L and H. L. collected sample; Y. T., H. L. and Z. L. dissected the tissue and sent the separated samples to the sequencing company; The genome assembly, annotation, and analysis are completed by D. L.; C-D. Z. provided guidance and suggestions for the writing and conceptualization of this article; The final manuscript has been read, edited, and approved by all authors.

Corresponding authors

Corresponding author: Zhonge Hou (houze@cnu.edu.cn)

References

1. Angela, M. G. et al. Evolutionary responses to warming. Trends in Ecology & Evolution, 36, 591–600, https://doi.org/10.1016/j.tree.2021.02.014 (2021).
2. Liu H. et al. Marine-montane transitions coupled with gill and genetic convergence in extant crustacean. Science Advances, 9, https://www.science.org/doi/10.1126/sciadv.adg4011 (2023).
3. Poynton, H. C. et al. The toxicogenome of *Hyalella azteca*: a model for sediment ecotoxicology and evolutionary toxicology. Environmental science & technology, 52, 6009–6022, 10.1021/acs.est.8b00837 (2018).

4. Harlıoğlu, M. M., & Farhadi, A. Importance of *Gammarus* in aquaculture. Aquaculture International, 26, 1327–1338, https://link.springer.com/article/10.1007/s10499-018-0287-6 (2018).
5. Averof, M. The crustacean *Parhyale*. Nature Methods, 19, 1015–1016, https://doi.org/10.1038/s41592-022-01596-y (2022).
6. Liu H. et al. Osmoregulatory evolution of gills promoted salinity adaptation following the sea-land transition of crustacean. Marine Life Science & Technology, 7, 205–217, https://doi.org/10.1007/s42995-025-00298-6 (2025).
7. Alther, R. et al. Optimizing laboratory cultures of *Gammarus fossarum* (Crustacea: Amphipoda) as a study organism in environmental sciences and ecotoxicology. Science of the Total Environment, 855, 158730, https://doi.org/10.1016/j.scitotenv.2022.158730 (2023).
8. Huang M. et al. Diversity of endemic cold-water amphipods threatened by climate warming in northwestern China. Diversity and Distributions, 30, e13798, https://doi.org/10.1111/ddi.13798 (2024).
9. Hou Z. et al. Past climate cooling promoted global dispersal of amphipods from Tian Shan montane lakes to circumboreal lakes. Global Change Biology, 28, 3830–3845, https://doi.org/10.1111/gcb.16160 (2022).
10. Schatz, M. C. et al. Assembly of large genomes using second-generation sequencing. Genome research, 20, 1165–1173, http://www.genome.org/cgi/doi/10.1101/gr.101360.109 (2010).
11. Tegenfeldt, F. et al. OrthoDB and BUSCO update: annotation of orthologs with wider sampling of genomes. Nucleic Acids Research, 53, D516–D522, https://doi.org/10.1093/nar/gkae987 (2025).
12. Zhang, H., Sun, S., Liu, J., Guo, Q., Meng, L., Chen, J., Xiang, X., Zhou, Y., Zhang, N., Liu, H., Liu, Y., Yan, G., Ji, Q., He, L., Cai, S., Cai, C., Huang, X., Xu, S., Xiao, Y., Zhang, Y., Wang, K., Liu, Y., Chen, H., Yue, Z., He, S., Wang, J., Yang, H., Liu, X., Seim, I., Gu, Y., Li, Q., Zhang, G., Lee, S.M.-Y., Kristiansen, K., Xu, X., Liu, S., and Fan, G. CNGB https://db.cngb.org/data_resources/assembly/CNA0142381
13. Liu, H., Zheng, Y., Zhu, B., Tong, Y., Xin, W., Yang, H., Jin, P., Hu, Y., Huang, M., Chang, W., Ballarin, F., Li, S., and Hou, Z. Genbank https://identifiers.org/ncbi/insdc.gca: GCA_030386875.1
14. Kao, D., Lai, A.G., Stamataki, E., Rosic, S., Konstantinides, N., Jarvis, E., Di Donfrancesco, A., Pouchkina-Stancheva, N., Semon, M., Grillo, M., Bruce, H., Kumar, S., Siwanowicz, I., Le, A., Lemire, A., Eisen, M.B., Extavour, C., Browne, W.E., Wolff, C., Averof, M., Patel, N.H., Sarkies, P., Pavlopoulos, A., and Aboobaker, A. Genbank https://identifiers.org/ncbi/insdc.gca: GCA_001587735.2
15. Poynton, H.C., Hasenbein, S., Benoit, J.B., Sepulveda, M.S., Poelchau, M.F., Hughes, D.S.T., Murali, S.C., Chen, S., Glastad, K.M., Goodisman, M.A.D., Werren, J.H., Vineis, J.H., Bowen, J.L., Friedrich, M., Jones, J., Robertson, H.M., Feyereisen, R., Mechler-Hickson, A., Mathers, N., Lee, C.E., Colbourne, J.K., Biales, A., Johnston, J.S., Wellborn, G.A., Rosendale, A.J., Cridge, A.G., Munoz-Torres, M.C., Bain, P.A., Manny, A.R., Major, K.M., Lambert, F.N., Vulpe, C.D., Tuck, P., Blalock, B.J., Lin, Y.Y., Smith, M.E., Ochoa-Acuna, H., Chen, M.M., Childers, C.P., Qu, J., Dugan, S., Lee, S.L., Chao, H., Dinh, H., Han, Y., Doddapaneni, H., Worley, K.C., Muzny, D.M., Gibbs, R.A., and Richards, S. Genbank https://identifiers.org/ncbi/insdc.gca: GCA_000764305.4
16. Patra, A.K., Chung, O., Yoo, J.Y., Kim, M.S., Yoon, M.G., Choi, J.H., and Yang, Y. Genbank https://identifiers.org/ncbi/insdc.gca: GCA_006783055.1
17. Cormier, A., Chebbi, M.A., Giraud, I., Wattier, R., Teixeira, M., Gilbert, C., Rigaud, T., and Cordaux, R. Genbank https://identifiers.org/ncbi/insdc.gca: GCA_016164225.1
18. Miron, W., and Pirro, S. Genbank https://identifiers.org/ncbi/insdc.gca: GCA_047292215.1
19. Edsinger, E., Kieras, M., and Pirro, S. Genbank https://identifiers.org/ncbi/insdc.gca: GCA_037179465.1

20. Liu, H. et al. Genomics of rafting crustaceans reveals adaptation to climate change in tropical oceans. Nat. Commun. https://doi.org/10.1038/s41467-026-69173-x (2026).
21. Direct Submission. Genbank https://identifiers.org/ncbi/insdc.gca: GCA_947561585.1
22. Nunez, J.C.B., Rand, D.M., Rong, S., Williams, S., Okami, N., Greenhill, M., Neil, K.B., Burley, J., Morgan, D.M., Ferranti, D.A., Brown, B.R., Lyons, A., and Spieler, A. Genbank https://identifiers.org/ncbi/insdc.gca: GCA_014899125.1
23. Rao, S. S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell, 159, 1665–1680, https://doi.org/10.1016/j.cell.2014.11.021 (2014).
24. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science, 326, 289–293, https://doi.org/10.1126/science.1181369 (2009).
25. Ranallo-Benavidez, T. R. et al. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nature Communications, 11, 1432, https://doi.org/10.1038/s41467-020-14998-3 (2020).
26. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics, 27, 764–770, https://doi.org/10.1093/bioinformatics/btr011 (2011).
27. Cheng H. et al. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nature Methods, 18, 170–175, https://doi.org/10.1038/s41592-020-01056-5 (2021).
28. Guan D. et al. Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics, 36, 2896–2898, https://doi.org/10.1093/bioinformatics/btaa025 (2020).
29. Heng Li & Richard Durbin. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics, 25, 1754–1760, https://doi.org/10.1093/bioinformatics/btp324 (2009).
30. Jiang H. et al. NextPolish: a fast and efficient genome polishing tool for long-read assembly. Bioinformatics, 36, 2253–2255, https://doi.org/10.1093/bioinformatics/btz891 (2020).
31. Heng Li, Richard Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics, 25, 1754–1760, https://doi.org/10.1093/bioinformatics/btp324 (2009).
32. Zeng X. et al. Chromosome-level scaffolding of haplotype-resolved assemblies using Hi-C data without reference genomes. Nature Plants, 10, 1184–1200, https://doi.org/10.1038/s41477-024-01755-3 (2024).
33. Zhou C. et al. YaHS: yet another Hi-C scaffolding tool. Bioinformatics, 39, btac808, https://doi.org/10.1093/bioinformatics/btac808 (2023).
34. Dudchenko, O. et al. The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. BioRxiv, 254797, https://www.biorxiv.org/content/10.1101/254797v1 (2018).
35. Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proceedings of the National Academy of Sciences, 117, 9451–9457, https://doi.org/10.1073/pnas.1921046117 (2020).
36. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. Current protocols in Bioinformatics, 5, 4–10, https://doi.org/10.1002/0471250953.bi0410s25 (2009).
37. Gabriel, L. et al. TSEBRA: transcript selector for BRAKER. BMC Bioinformatics, 22, 566, https://doi.org/10.1186/s12859-021-04482-0 (2021).
38. Jens, K. et al. Using intron position conservation for homology-based gene prediction. Nucleic Acids Research, 44, e89, https://doi.org/10.1093/nar/gkw092 (2016).
39. Haas, B.J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program

to Assemble Spliced Alignments. Genome biology, 9, R7, https://doi.org/10.1186/gb-2008-9-1-r7 (2008).
40. Shadab, A. et al. The UniProt website API: facilitating programmatic access to protein knowledge. Nucleic Acids Research, 53, W547–W553, https://doi.org/10.1093/nar/gkaf394 (2025).
41. Buchfink, B. et al. Fast and sensitive protein alignment using DIAMOND. Nature Methods, 12, 59–60, https://doi.org/10.1038/nmeth.3176 (2015).
42. Buchfink, B. et al. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nature Methods, 18, 366–368, https://doi.org/10.1038/s41592-021-01101-x (2021).
43. Philip, J. et al. InterProScan 5: genome-scale protein function classification. Bioinformatics, 30, 1236–1240, https://doi.org/10.1093/bioinformatics/btu031 (2014).
44. Chen C. et al. TBtools: An Integrative Toolkit Developed for Interactive Analyses of Big Biological Data. Molecular Plant, 13, 1194–1202, https://doi.org/10.1016/j.molp.2020.06.009 (2020).
45. ENA Sequence Read Archive https://www.ebi.ac.uk/ena/browser/view/ERR16720729 (2026).
46. ENA Sequence Read Archive https://www.ebi.ac.uk/ena/browser/view/ERR16722743 (2026).
47. ENA Sequence Read Archive https://www.ebi.ac.uk/ena/browser/view/ERR16729500 (2026).
48. NCBI GenBank https://identifiers.org/insdc.gca:GCA_980912865.2.
49. Lu D. CNGBdb. https://db.cngb.org/data_resources/project/CNP0008126 (2026).
50. Lu, D. Functional annotation of gene. figshare. Figure. https://doi.org/10.6084/m9.figshare.31698019.v1 (2026).
51. Schultz, D.T. et al. Ancient gene linkages support ctenophores as sister to other animals. Nature, 618, 110–117, https://doi.org/10.1038/s41586-023-05936-6 (2023).

Ethics declarations

Competing interests

The authors declare no competing interests.