scientific data

OPEN

DATA DESCRIPTOR

A high-quality chromosome-level genome assembly and annotation of the Amur grayling (*Thymallus grubii*)

Gaochao Wang[1], Kai Ma[1], Tianqing Huang[1], Enhui Liu[1], Wei Gu[1], Peng Fan[1], Kaibo Ge[1], Yunchao Sun[1], Datian Li[1], Junwei Wang[2], Xinyang Wang[3], Gefeng Xu[1] ✉ & Yongquan Zhang[1] ✉

Graylings (*Thymallus*) represent one of the earliest diverged salmonid subfamilies and provide crucial insights into the evolutionary consequences of the salmonid-specific whole genome duplication event (Ss4R). Here, we present the first chromosome-level genome assembly of the Amur grayling (*Thymallus grubii*), an economically and ecologically important species endemic to East Asian freshwater systems. The final assembly spans 1,754.51 Mb with 46 pseudo-chromosomes, achieving a contig N50 of 3.03 Mb and anchoring 99.97% of sequences to chromosomes. The genome contains 31,867 protein-coding genes with 97.58% functionally annotated. Repetitive sequences comprise 52.64% of the genome, dominated by DNA transposons (22.06%). BUSCO analysis indicates 97.9% completeness for conserved Actinopterygii genes. This high-quality genome assembly provides an essential resource for salmonid comparative genomics, aquaculture breeding programs, and conservation genetics of this economically important species.

Background & Summary

The Amur grayling (*Thymallus grubii*), also referred to as Heilongjiang grayling, is endemic to the Amur River basin in northeastern Asia and represents one of the most phylogenetically significant yet genomically understudied members of the Salmonidae family[1,2]. As part of the Thymallinae subfamily, this species occupies a crucial position for understanding early salmonid evolution, particularly the genomic consequences of the salmonid-specific whole genome duplication (Ss4R) that occurred 80–100 million years ago[3,4].

From a genomic evolution perspective, graylings are particularly fascinating because they have retained ancestral chromosome characteristics following the Ss4R event. Unlike other salmonids that underwent extensive chromosomal fusions (such as Atlantic salmon with $2n = 58$), graylings maintained high chromosome numbers ($2n = 92–102$), suggesting alternative post-duplication evolutionary pathways[5–8]. The European grayling genome revealed fusion-free chromosomes dominated by inversions rather than fusions[8], but whether this pattern extends to the Amur grayling genome remained unknown.

The economic importance of *T. grubii* extends across multiple sectors, supporting significant recreational fisheries in its native range and representing a cultural icon in northeastern Asian aquatic ecosystems[9]. The species' cold-water adaptation, rapid growth rates, and high-quality flesh make it an attractive candidate for aquaculture development[10,11]. Recent studies have demonstrated its potential for intensive aquaculture, with research focusing on optimal dietary conditions and growth performance parameters[11,12]. However, conservation concerns are mounting due to habitat degradation and climate-induced changes in the Amur River system. The species' restricted distribution makes it particularly vulnerable to environmental changes, while population genetic studies have revealed significant phylogeographic structure within the Amur River basin[13,14].

[1]State Key Laboratory of Mariculture Biobreeding and Sustainable Goods, National and Local Joint Engineering Laboratory for Freshwater Fish Breeding, Heilongjiang River Fisheries Research Institute, Chinese Academy of Fishery Sciences, Harbin, 150070, China. [2]Shandong Wanzefeng Ocean Development Group Co., Ltd., Rizhao, 276800, China. [3]Heilongjiang Agricultural Engineering Vocational College, Harbin, 150025, China. ✉e-mail: xugefeng@hrfri.ac.cn; zyq0805@163.com

Despite the evolutionary and practical importance of the Amur grayling, genomic resources remain scarce compared to other salmonids. While chromosome-level assemblies exist for Atlantic salmon[3], European grayling[8], and other commercially important species[15–20], no high-quality genome assembly has been available for this species. This genomic gap has limited comparative studies of salmonid chromosome evolution and hindered the development of molecular tools for *T. grubii* research and management.

Here, we present the first high-quality chromosome-level genome assembly for *T. grubii*, employing PacBio HiFi long-read sequencing, Hi-C proximity ligation, and comprehensive transcriptome data. This 1.75 Gb assembly with 46 pseudo-chromosomes provides unprecedented insights into grayling genome organization, enables comparative analysis of salmonid chromosome evolution, and establishes essential genomic infrastructure for breeding programs and conservation genetics applications.

Methods

Sample collection and sequencing. A female *T. grubii* individual was collected at the Bohai Cold Water Fish Test Station, Heilongjiang Fisheries Research Institute, Chinese Academy of Fishery Sciences. The individual was anaesthetized and tissues including muscle, heart, liver, kidney and spleen were obtained. Tissues were rapidly preserved in liquid nitrogen and stored at −80 °C until further analysis. The fish was handled in accordance with the guidelines for the Care, Ethics, and Safety Inspection of Laboratory Animals of the Heilongjiang Fisheries Research Institute of the Chinese Academy of Fishery Sciences. Ethical approval was obtained from the Institutional Review Board.

High molecular weight genomic DNA was isolated from muscle tissue using the MagAttract HMW DNA Kit (Qiagen, Germantown, MD, USA) following manufacturer's guidelines. For PacBio HiFi sequencing, DNA was fragmented to 15–20 kb using the MegaRuptor 3 (Diagenode, Denville, NJ, USA). Library preparation was conducted using the SMRTbell Express Template Prep Kit 2.0, followed by treatment with the Enzyme Clean Up Kit (Pacific Biosciences, Menlo Park, CA, USA). Sequencing was performed on the PacBio Revio platform, generating 89.59 Gb of HiFi reads. Raw data were processed using the CCS program v4.2.0 (https://ccs.how/) with parameters: minimum pass of 3, minimum subread length of 50 bp, and minimum predicted accuracy of 0.99.

For Hi-C sequencing, liver tissue was fixed in formaldehyde, lysed, and cross-linked DNA was digested overnight with DpnII. Sticky ends were biotinylated and proximity-ligated to create chimeric junctions, which were enriched and fragmented to 300–500 bp. Sequencing was performed on the DNBSEQ-T7 system (MGI, Shenzhen, China), yielding 224.09 Gb of 150 bp paired-end reads. Additionally, whole genome shotgun sequencing using MGI PE150 technology generated 110.48 Gb of paired-end genomic reads from muscle tissue for a pre-assembly genome survey aimed at estimating genome size, heterozygosity, and repeat content.

For transcriptome sequencing, total RNA was extracted from four tissues (including muscle, heart, kidney and spleen) using TRIzol reagent (Invitrogen, CA, USA) and purified using the RNeasy Mini Kit (Qiagen, CA, USA). RNA quality was assessed using agarose gel electrophoresis and Agilent Bioanalyzer 2100 system, with all samples achieving RNA Integrity Number (RIN) values above 8.5. mRNA was enriched using Oligo (dT) beads. Sequencing libraries were prepared using the MGIEasy RNA Library Prep Set (MGI, Shenzhen, China) and sequenced on the DNBSEQ-T7 platform, generating 150 bp paired-end reads. RNA-seq data were successfully generated from four tissues with varying yields: muscle (9.54 Gb), heart (11.29 Gb), kidney (9.42 Gb), and spleen (12.27 Gb). Mapping rates to the assembled genome ranged from 88.42% (heart) to 92.60% (muscle), providing extensive transcript evidence for gene structure prediction and validation.

Genome survey analysis. A genome survey using GenomeScope2.0[21] based on k-mer analysis estimated the genome size at 1,654.00 Mb, with a low heterozygosity of 0.303%. The unimodal k-mer distribution confirmed its diploid nature (Fig. 1). While the estimated genome size is comparable to the final assembly size (1,754.51 Mb), the k-mer-based repeat estimate (30.6%) was substantially lower than the final annotation result (52.64%). This is expected, as k-mer methods are less sensitive to older, diverged repeats that are effectively identified by homology-based annotation. The slight discrepancy in genome size is typical in such projects, particularly for species with high repeat content like salmonids. K-mer-based methods often underestimate size due to their challenges in accurately modeling highly repetitive or divergent regions, such as transposable elements from the Ss4R event, which can lead to collapsed k-mer counts. In contrast, the final assembly benefits from long-read PacBio HiFi sequencing and Hi-C scaffolding, which better resolve these complex regions and yield a more complete genome representation. The survey thus provided a solid foundation by confirming diploidy and offering a reasonable genome size estimate.

Genome assembly and scaffolding. Initial contig assembly was performed using HiFiasm v0.19.8-r603[22] with default parameters, utilizing the PacBio HiFi reads. Subsequently, purge_dups v1.2.5[23] was employed to eliminate redundant haplotigs in the primary assemblies. Hi-C scaffolding was conducted using the ALLHIC scaffolding pipeline[24]. To guide the scaffolding process, the European grayling (NCBI accession: GCA_023634145.1, *T. thymallus*) was selected as the reference due to its close phylogenetic proximity to *T. grubii* within the Thymallinae subfamily and shared karyotypic characteristics, including high chromosome numbers and limited fusions following the Ss4R event. These similarities facilitate accurate allelic contig grouping and reduce scaffolding errors compared to more divergent salmonids. Initially, GMAP[25] was used to align coding sequences of reference species to generate an allelic contig table to filter out noisy Hi-C signals. The contigs were divided into different homologous groups based on the allelic contig table to reduce scaffolding complexity. Finally, unscaffolded contigs were globally rescued using ALLHiC_rescue. The scaffolding results were manually reviewed using Juicebox[26]. The Hi-C interaction heatmap was generated with HiCExplorer[27].

The mitochondrial genome was assembled using mitohifi v3.2.1[28]. Contigs from the HiFiasm assembly that aligned to mitochondrial sequences over 50% of their length were removed from the final assembly.

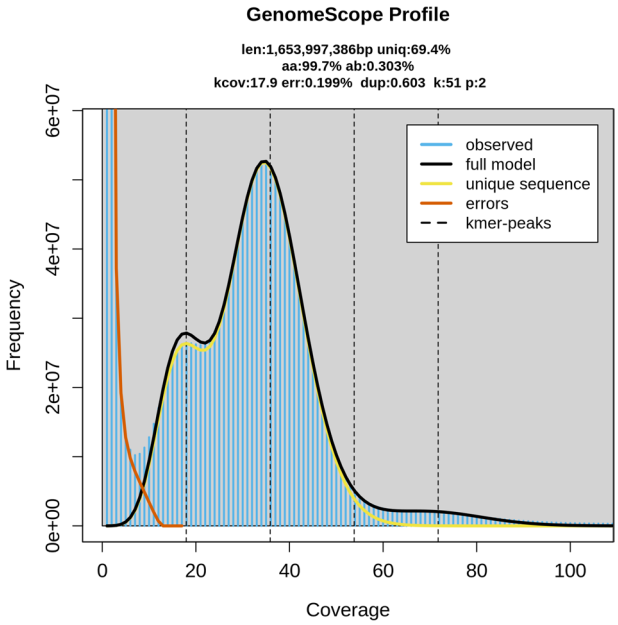

Fig. 1 Genome survey analysis of *T. grubii*. K-mer frequency distribution plots for *T. grubii*. The x-axis shows K-mer depth, and the y-axis represents the number of K-mers.

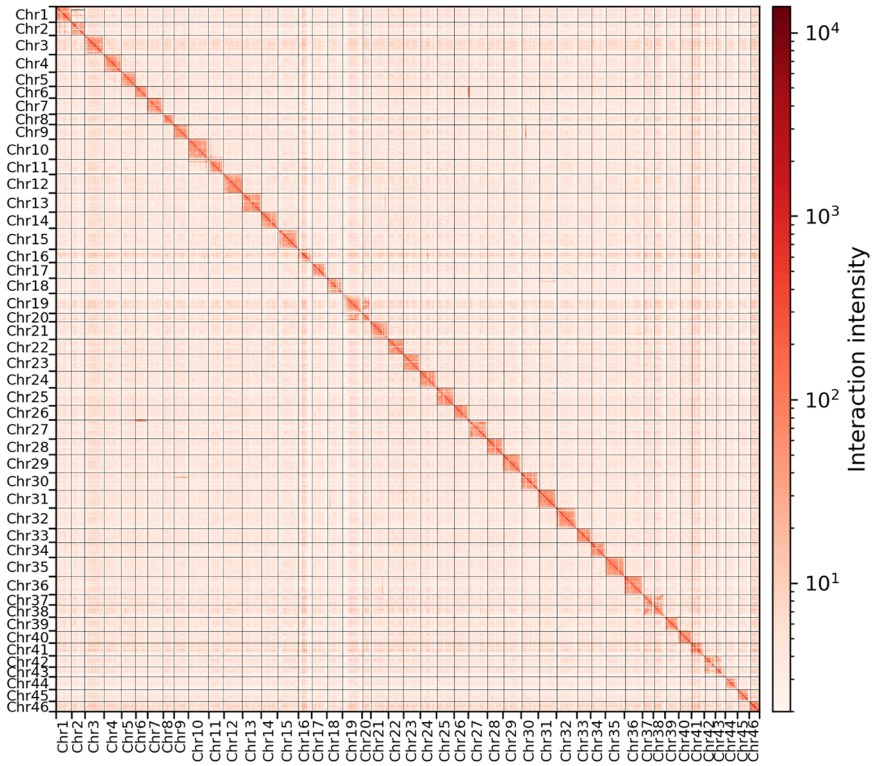

Fig. 2 Hi-C contact map of *T. grubii* genome assembly. The x and y axes represent chromosomes, with color intensity indicating interaction frequency between genomic regions. Chromosome boundaries are indicated by black lines for clarity.

Hi-C contact maps reveal clear chromosomal territories with strong intra-chromosomal interactions and minimal inter-chromosomal contacts (Fig. 2). The contact patterns support the chromosome-level organization and validate the scaffolding accuracy. No evidence of large-scale misassemblies is apparent in the Hi-C interaction patterns. The genomic characteristics of the assembled genome are visualized in Circos plots showing GC content, gene density, repeat sequence density, and chromosome synteny blocks (Fig. 3).

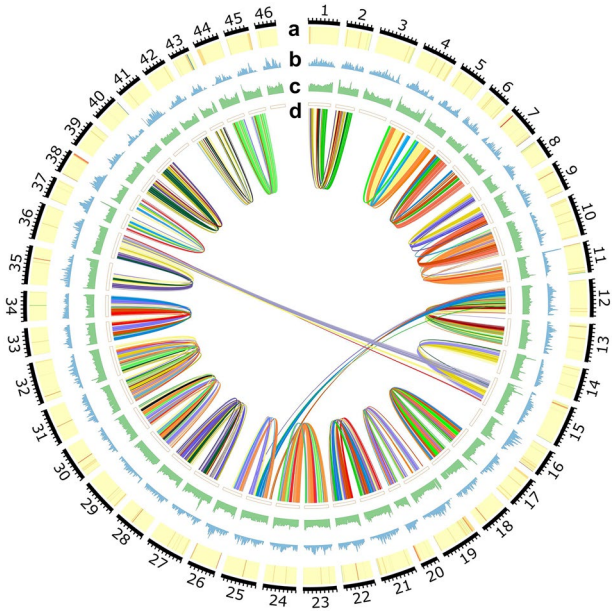

Fig. 3 Circos plots of the genomic characteristics of *T. grubii* genome assembly. Circos plots illustrating the chromosome-level genome assemblies of *T. grubii*. The plot displays the following layers from the outermost to the innermost circle: (a) a heatmap of GC content; (b) a gene density plot; (c) a repeat sequence density plot; (d) chromosome synteny blocks (showing blocks >1 Mb).

The final chromosome-level assembly spans 1,754.51 Mb, comprising 46 pseudo-chromosomes plus one mitochondrial genome (ChrM) (Table 1). The assembly demonstrates high contiguity with a contig N50 of 3.03 Mb and scaffold N50 of 39.65 Mb. A total of 99.97% of assembled sequences are anchored to chromosomes, indicating successful chromosome-scale scaffolding. The largest contig reaches 14.83 Mb, while the largest scaffold spans 51.7 Mb (Table 2).

Repetitive element annotation. Repetitive sequences were identified using RepeatMasker v4.1.6[29] against the RepBase database[30], supplemented with de novo repeat identification using RepeatModeler v2.0.5[31]. This process involved de novo identification of transposable elements with RepeatScout v1.0.6[32] and RECON v1.08[33]. LTRharvest[34] and LTR_retriever v2.9.5[35] were employed to identify high-quality long terminal repeat (LTR) families. CD-HIT v4.8.1[36] was used to remove redundant TEs from the library. The identified TE families were annotated and classified using the Dfam v3.7[37] and Repbase v20181026[30] databases. Tandem repeats were identified using Tandem Repeats Finder[38].

Repetitive sequences comprise 923.51 Mb (52.64%) of the *T. grubii* genome (Table 3). This repetitive content is consistent with other salmonid genomes and reflects the genomic consequences of the Ss4R whole genome duplication event[3,4]. DNA transposons represent the largest category at 387.06 Mb (22.06%), followed by LINEs at 195.06 Mb (11.12%) and LTR elements at 88.36 Mb (5.04%). Among LTR retrotransposons, Ty1/Copia elements (39.87 Mb, 2.27%) and Gypsy/DIRS1 elements (28.71 Mb, 1.64%) were most abundant. SINEs contributed 41.89 Mb (2.39%) to the repetitive landscape. Tandem repeats totaled 76.63 Mb (4.36%). The high proportion of interspersed repeats (48.82%) reflects the transposon-mediated repeat expansion characteristic of salmonid genomes following whole genome duplication.

Gene structure prediction and functional annotation. Gene structure prediction was conducted using the ETP model of the BRAKER3 pipeline v3.0.8[39,40], which integrates homologous protein-based evidence and transcriptome data. Initially, ProtHint[41] was utilized to align 9,805,833 Vertebrata orthologous genes, obtained from OrthoDB v11[42], to the genome sequences after duplication masking. Concurrently, RNA-Seq data were aligned to the duplicate-masked genome using HISAT2 v2.1.0[43]. GeneMark-ETP+[44] then used evidence from these sources to make initial unsupervised gene predictions. Subsequently, AUGUSTUS v3.5.0[45] was trained for two rounds using the initial high-quality gene models. Final gene structures were predicted with AUGUSTUS, combining transcriptome and homologous protein evidence. UTRs were predicted using GUSHR v1.0.0[46].

The comprehensive annotation pipeline identified 31,867 protein-coding genes generating 41,902 transcripts (Table 4). Genes average 10,407.56 bp in length with an average of 12.32 exons per gene. The total exonic sequence comprises 68.08 Mb, representing approximately 3.9% of the genome. Total gene length including introns spans 331.66 Mb (18.9% of the genome), indicating extensive intergenic regions typical of salmonid genomes.

Protein-coding genes were functionally annotated using multiple databases including NR (RefSeq non-redundant proteins, 2020) and Swiss-Prot[47] using DIAMOND v2.0.9.147[48] with parameters:–more-sensitive -p 64 -e 1e-6–max-hsps 1 -k 1 -f 6. The eggNOG-mapper v2.1.12[49] annotation pipeline was used

Chromosome	Length (bp)	Contigs	Chromosome	Length (bp)	Contigs
Chr1	38,616,233	50	Chr24	40,546,842	46
Chr2	32,855,317	51	Chr25	44,120,899	35
Chr3	48,338,156	46	Chr26	34,924,661	46
Chr4	43,812,173	33	Chr27	47,587,119	40
Chr5	34,903,637	25	Chr28	39,653,670	31
Chr6	29,441,728	50	Chr29	44,584,444	23
Chr7	39,425,273	40	Chr30	42,328,509	39
Chr8	26,202,112	22	Chr31	45,329,829	20
Chr9	36,244,024	37	Chr32	51,701,102	34
Chr10	50,575,919	36	Chr33	34,270,833	38
Chr11	37,858,395	30	Chr34	36,981,557	42
Chr12	46,273,032	34	Chr35	47,492,427	25
Chr13	46,729,120	30	Chr36	45,131,135	29
Chr14	41,002,823	44	Chr37	26,311,800	38
Chr15	51,656,326	42	Chr38	29,322,451	25
Chr16	33,434,292	48	Chr39	34,125,346	42
Chr17	38,212,852	33	Chr40	31,507,621	19
Chr18	38,028,410	52	Chr41	31,334,277	63
Chr19	49,616,253	62	Chr42	28,262,599	52
Chr20	21,343,220	36	Chr43	24,473,442	45
Chr21	43,611,012	43	Chr44	30,722,859	51
Chr22	38,439,102	22	Chr45	29,864,843	51
Chr23	41,255,341	28	Chr46	25,485,447	62
Unanchored contig1	489804	1	Unanchored contig3	15161	1
Unanchored contig2	27846	1	Unanchored contig4	21305	1
ChrM	16,659	1	Total	1,754,509,237	1,795

Table 1. Statistics of pseudo-chromosomes of *T. grubii* genome.

Assembly metric	Number
Genome Assembly size (Mb)	1754.51
Nuclear pseudo-chromosomes count	46
Mitochondrial genome	1
Contig count	1795
Largest contig size (Mb)	14.83
Contig N50 length (Mb)	3.03
Contig L50	152
Ratio of length of contigs anchored to chromosomes	0.9997
Unanchored contigs	4
Scaffold N50 (Mb)	39.65
Repetitive sequences (length Mb/percent %)	923.51 Mb/52.64%
Protein-coding genes/transcripts	31867/41902
GC content (%)	42.89
Complete BUSCOs (genome %)	97.9

Table 2. Summary of genome assembly statistics.

to annotate genes to the eggNOG 5.0 database[50]. This process also provided Gene Ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway annotations. Additionally, InterProScan v5.55[51] was employed for annotation against the PFAM database[52].

Functional annotation achieved high coverage across multiple databases (Table 5), with 31,096 genes (97.58%) receiving functional assignments. The eggNOG database provided annotations for 29,207 genes, while NR database covered 31,058 genes. Gene Ontology terms were assigned to 21,692 genes, and 26,524 genes matched entries in the Swiss-Prot database. Only 770 genes (2.42%) remained without functional annotation, likely representing species-specific or rapidly evolving gene families.

Syntenic analysis. To identify syntenic blocks in the assembled genome, all-vs-all BLAST of primary protein sequences was performed using DIAMOND with parameters–threads 80–evalue 1e-6–more-sensitive–outfmt 6.

Repeat type	Number	Length (bp)	Proportion in genome (%)
SINEs	322,586	41,891,552	2.39
LINEs	564,591	195,055,563	11.12
LTR elements	219,867	88,359,443	5.04
LTR-Ty1/Copia	77,697	39,872,971	2.27
LTR-Gypsy/DIRS1	71,547	28,711,479	1.64
DNA transposons	1,405,111	387,064,292	22.06
Unclassified	850,813	144,227,726	8.22
Total interspersed repeats (TEs)		856,598,576	48.82
Rolling-circles	3,007	927,675	0.05
Small RNA	63,105	14,050,312	0.8
Satellites	11,854	2,522,280	0.14
Simple repeats	605,258	53,593,055	3.05
Low complexity	68,201	6,462,852	0.37
Total Tandem repeats	748,418	76,628,499	4.36
Total masked		923,509,832	52.64

Table 3. Repetitive element composition in the genome.

Annotation metric	Number
Total length of protein-coding gene (bp)	331,657,646
Total exons length (bp)	68,076,439
Protein-coding genes count	31,867
Protein-coding transcripts count	41,902
Exon count	392,521
Average length of protein-coding gene (bp)	10,407.56
Average length of exons	173.43
Average Number of exons per gene	12.32

Table 4. Genome annotation statistics.

Database	Number
eggNOG	29,207
KOG	10,679
GO	21,692
NR	31,058
Swissprot	26,524
Pfam	26,524
Interproscan	28,574
Total of Annotated	31,096
Total of Unnotated	770
Ratio of Annotated	97.58%

Table 5. Functional annotation results.

Collinear blocks were then identified using MCScanX[53]. For pairwise synteny visualization, the Python version of MCScan, implemented in jcvi[54], was employed to align CDS sequences and generate synteny plots.

Comparative analysis with Northern pike (NCBI accession: GCF_011004845.1, *Esox lucius*) as a pre-duplication outgroup reveals expected salmonid chromosome evolution patterns (Fig. 4). The synteny blocks demonstrate extensive conservation between Northern pike chromosomes and *T. grubii* chromosomes, with most pike chromosomes showing conserved syntenic relationships with multiple *T. grubii* chromosomes. The connecting lines in the synteny plot illustrate numerous collinear blocks that span substantial chromosomal regions, indicating remarkable conservation of ancestral gene order despite ~80–100 million years of divergence since the Ss4R event[3,4]. The analysis confirms that *T. grubii* has retained near-ancestral chromosome numbers through inversions and intrachromosomal rearrangements rather than the extensive fusion events that characterize other salmonid lineages, consistent with previous findings in European grayling[8]. The comprehensive genome-wide synteny coverage suggests graylings represent a unique evolutionary solution to genomic instability following whole genome duplication.

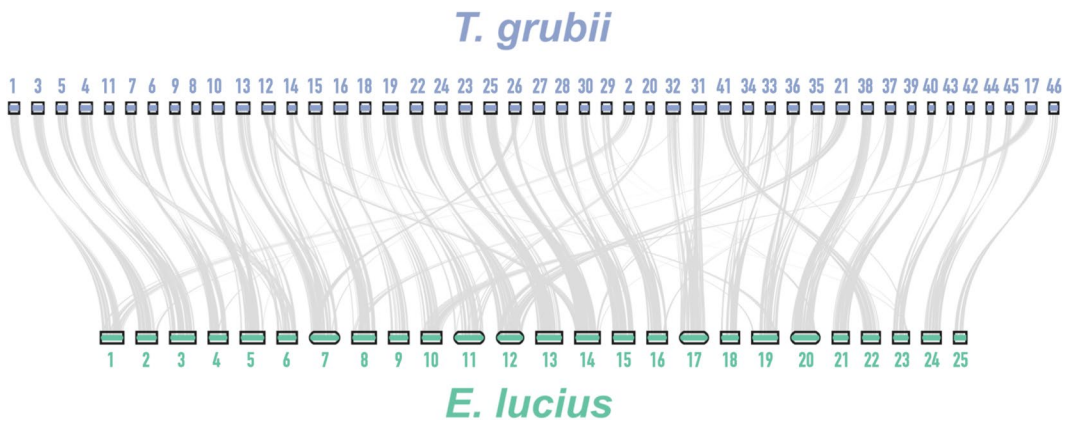

Fig. 4 Synteny analysis between *T. grubii* and *E. lucius*. Connecting lines show conserved syntenic blocks across substantial chromosomal regions.

Data Records

All raw sequencing data supporting this study are publicly available and deposited in the NCBI Sequence Read Archive (SRA) under accession SRP601401[55] (BioProject PRJNA1292408). This includes PacBio HiFi long reads (89.6 Gb), MGI paired-end genomic reads (110.5 Gb), Hi-C proximity ligation reads (224.1 Gb), and RNA-seq data from four tissues (Total 42.5 Gb: muscle, heart, kidney, spleen). Data are organized by experiment type in the SRA submission, with metadata including sample details, read lengths, and quality scores. The final chromosome-level genome assembly (FASTA format) was deposited in NCBI GenBank under accession GCA_051833735.1[56]. Genome annotation files (GTF format for genes/transcripts, plus cds and protein sequences in FASTA) were deposited in the Figshare database[57].

Technical Validation

Assembly quality was evaluated using multiple metrics. The assembly achieved high contiguity with contig N50 of 3.03 Mb, scaffold N50 of 39.65 Mb, and 99.97% of sequences anchored to 46 chromosomes with only 1,744 gaps across 1,795 scaffolds (Table 2). BUSCO[58] analysis indicates 97.9% completeness against the Actinopterygii gene set, demonstrating high assembly quality and gene space coverage (Table 2).

Hi-C contact maps show clear chromosomal territories with strong intra-chromosomal interactions and minimal inter-chromosomal contacts, validating chromosome-level scaffolding accuracy (Fig. 2).

Data availability

The datasets generated and analyzed in this study are publicly available in NCBI Sequence Read Archive (SRA), NCBI GenBank, and Figshare repositories. For detailed access instructions, including file lists, see the Data Records section.

Code availability

Bioinformatics analyses in this study were conducted following the official manuals and recommended protocols provided by the software developers. Details regarding the software-specific parameters employed are outlined in the Methods section.

Received: 7 August 2025; Accepted: 31 October 2025;
Published online: 18 December 2025

References

1. Weiss, S. J. *et al.* Global systematic diversity, range distributions, conservation and taxonomic assessments of graylings (Teleostei: Salmonidae; *Thymallus* spp.). *Org. Divers. Evol.* **21**, 25–42 (2021).
2. Macqueen, D. J. & Johnston, I. A. A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. *Proc. R. Soc. B* **281**, 20132881 (2014).
3. Lien, S. *et al.* The Atlantic salmon genome provides insights into rediploidization. *Nature* **533**, 200–205 (2016).
4. Gundappa, M. K. *et al.* Genome-wide reconstruction of rediploidization following autopolyploidization across one hundred million years of salmonid evolution. *Mol. Biol. Evol.* **39**, msab310 (2022).
5. Phillips, R. & Ráb, P. Chromosome evolution in the Salmonidae (Pisces): an update. *Biol. Rev.* **76**, 1–25 (2001).
6. Robertson, F. M. *et al.* Lineage-specific rediploidization is a mechanism to explain time-lags between genome duplication and evolutionary diversification. *Genome Biol.* **18**, 111 (2017).
7. Crête-Lafrenière, A., Weir, L. K. & Bernatchez, L. Framing the Salmonidae family phylogenetic portrait: a more complete picture from increased taxon sampling. *PLoS One* **7**, e46662 (2012).
8. Sävilammi, T. *et al.* The chromosome-level genome assembly of European grayling reveals aspects of a unique genome evolution process within salmonids. *G3 Genes Genomes Genet.* **9**, 1283–1294 (2019).
9. Weiss, S. *et al.* Multiple species of grayling (*Thymallus* sp.) found in sympatry in remote tributary of Amur River. *Zool. Scr.* **49**, 117–128 (2020).
10. Han, H. *et al.* Nutritional quality of muscle and eggs in three species of wild female graylings from Heilongjiang river. *Journal of Guangdong Ocean University* **44**, 55–62 (2024).

11. Lu, S. *et al*. Effects of dietary fish oil levels on growth performance, lipid metabolism, hepatic health, nonspecific immune response, and intestinal microbial community of juvenile Amur grayling (*Thymallus grubii*). *Aquac. Nutr.* **2024**, 8587410 (2024).
12. Fan, Z. *et al*. Liver transcriptome and physiological analyses preliminarily revealed the adaptation mechanisms of Amur grayling (*Thymallus arcticus grubei*, Dybowski, 1869) fry for dietary lipid nutrition. *Front. Vet. Sci.* **11**, 1369845 (2024).
13. Ma, B. *et al*. Phylogeography and population genetic structure of Amur grayling *Thymallus grubii* in the Amur basin. *Asian-Australas. J. Anim. Sci.* **25**, 935 (2012).
14. Froufe, E. *et al*. Identification of reproductively isolated lineages of Amur grayling (*Thymallus grubii* Dybowski 1869): concordance between phenotypic and genetic variation. *Mol. Ecol.* **12**, 2345–2355 (2003).
15. Zhu, W. *et al*. A chromosome-level genome of *Brachymystax tsinlingensis* provides resources and insights into salmonids evolution. *G3 Genes Genomes Genet.* **12**, jkac162 (2022).
16. Gao, G. *et al*. The generation of the first chromosome-level de novo genome assembly and the development and validation of a 50K SNP array for the St. John River aquaculture strain of North American Atlantic salmon. *G3 Genes Genomes Genet.* **13**, jkad138 (2023).
17. Christensen, K. A. *et al*. Chinook salmon (*Oncorhynchus tshawytscha*) genome and transcriptome. *PLoS One* **13**, e0195461 (2018).
18. Rondeau, E. B. *et al*. Insights from a chum salmon (*Oncorhynchus keta*) genome assembly regarding whole-genome duplication and nucleotide variation influencing gene function. *G3 Genes Genomes Genet.* **13**, jkad127 (2023).
19. Hansen, T. *et al*. The genome sequence of the brown trout, *Salmo trutta* Linnaeus 1758. *Wellcome Open Res.* **6**, 108 (2021).
20. Smith, S. R. *et al*. A chromosome-anchored genome assembly for Lake Trout (*Salvelinus namaycush*). *Mol. Ecol. Resour.* **22**, 679–694 (2022).
21. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1432 (2020).
22. Cheng, H. *et al*. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
23. Guan, D. *et al*. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).
24. Zhang, X. *et al*. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat. Plants* **5**, 833–845 (2019).
25. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
26. Durand, N. C. *et al*. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
27. Wolff, J. *et al*. Galaxy HiCExplorer 3: a web server for reproducible Hi-C, capture Hi-C and single-cell Hi-C data analysis, quality control and visualization. *Nucleic Acids Res.* **48**, 177–184 (2020).
28. Uliano-Silva, M. *et al*. MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads. *BMC Bioinformatics* **24**, 288 (2023).
29. RepeatMasker v4.1.6 link, http://www.repeatmasker.org.
30. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).
31. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA* **117**, 9451–9457 (2020).
32. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics* **21**, i351–i358 (2005).
33. Bao, Z. & Eddy, S. R. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res.* **12**, 1269–1276 (2002).
34. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).
35. 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).
36. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
37. Hubley, R. *et al*. The Dfam database of repetitive DNA families. *Nucleic Acids Res.* **44**, D81–D89 (2016).
38. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
39. Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M. & Stanke, M. BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* **32**, 767–769 (2016).
40. Brůna, T. *et al*. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom. Bioinform.* **3**, lqaa108 (2021).
41. Brůna, T. *et al*. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genom. Bioinform.* **2**, lqaa026 (2020).
42. Kriventseva, E. V. *et al*. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* **47**, D807–D811 (2019).
43. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
44. Lomsadze, A., Burns, P. D. & Borodovsky, M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res.* **42**, e119 (2014).
45. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644 (2008).
46. GUSHR v1.0.0 link, https://github.com/Gaius-Augustus/GUSHR.
47. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45–48 (2000).
48. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
49. Huerta-Cepas, J. *et al*. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
50. Huerta-Cepas, J. *et al*. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).
51. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
52. Mistry, J. *et al*. Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
53. Wang, Y. *et al*. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
54. jcvi link, https://doi.org/10.5281/zenodo.31631.
55. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP601401 (2025).
56. Wang, G. C. *Genbank* https://identifiers.org/insdc.gca:GCA_051833735.1 (2025).
57. Wang, G. C. Amur grayling genome files. *Figshare* https://doi.org/10.6084/m9.figshare.29819129 (2025).
58. Simão, F. A. *et al*. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).

Acknowledgements
Funding for this research was provided by the Central Public-interest Scientific Institution Basal Research Fund, HRFRI, (No. HSY202513Q), China Agriculture Research System of MOF and MARA (CARS-46), Identification and evaluation of characteristic Aquatic Animal Germplasm Resources and creation of improved varieties in Heilongjiang Province, Technology Collaborative Innovation and Extension System for the Freshwater Fish Industry of Heilongjiang Province, the earmarked fund for Xinjiang Agriculture Research System (XJARS-08-03), Central Government Guidance Funds for Local Science and Technology Development (ZY24QY03).

Author contributions
Gaochao Wang designed the study, analyzed the data, and wrote the original draft. Kai Ma, Tianqing Huang, Enhui Liu, Wei Gu, Peng Fan, Kaibo Ge, Yunchao Sun, Datian Li, Junwei Wang and Xinyang Wang contributed to specimen collection and experiments. Gefeng Xu and Yongquan Zhang supervised the project, and serve as corresponding authors. All authors contributed to manuscript preparation, and approved the final manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to G.X. or Y.Z.

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

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